C

Cadillac Tax: An Offset to the Tax Subsidy for Employer-Sponsored Health Insurance

Jane G. Gravelle
Congressional Research Service, Washington, DC, USA

Abstract

This entry will discuss the Cadillac tax imposed on high-cost employer health insurance (in excess of a dollar floor), its objectives as part of the Affordable Care Act (health reform) in 2010, and its subsequent modification in 2015. The history of the Act suggests that it was an alternative to limiting the exclusion of health insurance compensation from payroll and income taxes of employees to a dollar cap. The objective of the tax, in addition to providing financing for health reform, was to discourage the provision of high-cost insurance coverage and reduce health care spending. These effects, in turn, depend on the magnitude of the tax, the scope of coverage over time (which is expected to increase) and how the tax compares to the value of the exclusion of the benefits to employees from income tax and payroll taxes. Although the tax is imposed on insurers, its burden is expected to be passed on to labor income. Taxes will rise whether firms keep the high-cost insurance (and pay tax directly) or reduce insurance coverage and substitute taxable wages. The tax, currently imposed at 40%, is imposed on a tax-exclusive basis (as a percentage of the cost before taxes) while tax rates on employee wages are imposed on a tax-inclusive basis (as a percentage of cost inclusive of wages). For some employees the tax burden is smaller if the Cadillac tax is retained, while for others it is better to reduce insurance coverage. Those instances in which retaining the Cadillac tax is better have increased with the provision making the tax deductible from income tax, adopted in 2015. Only in circumstances where coverage is reduced will potential reductions in health care spending and efficiency gains from reductions in moral hazard be realized.

Keywords

Affordable Care Act; Cadillac tax; Employer health exclusion; Excise tax; Health care market; High-cost health insurance; Tax incidence; Tax-exclusive rate; Tax-inclusive rate

JEL Classification

H22; H25; I18

The views in this study do not reflect the views of the Congressional Research Service.

Macmillan Publishers Ltd (ed.), *The New Palgrave Dictionary of Economics*,
https://doi.org/10.1057/978-1-349-95189-5

The Cadillac tax is a US excise tax on high-cost employer-provided or union-provided health insurance plans that imposes a tax on the excess cost above a dollar amount with the intent of reducing the generosity of health plan coverage and health care expenditures. It was first included in the Affordable Care Act and was set to commence 6 years after the ACA went into effect in 2018. Under this plan the tax is imposed at a 40 percent rate on the excess cost above a dollar threshold adjusted for the health cost adjustment percentage. The Consolidated Appropriations Act 2016 (enacted at the end of 2015) delayed the implementation to 2020. Were the tax to have been in effect in 2018, it would have applied to health insurance costs in excess of $10,200 for single coverage and $27,500 for non-single coverage. The Consolidated Appropriations Act provided that the 2018 amounts be adjusted by the consumer price index plus 1% for 2019, and for the consumer price index in each following year. Assuming a CPI-U of 2.3% per year, these amounts will be $10,800 for single (self-only) coverage and $29,100 for family coverage in 2020.

The tax base includes both employer and employee contributions to insurance premiums, various health savings or flexible spending accounts (flexible spending accounts, health savings accounts, health reimbursement accounts and medical savings accounts), self-employed plans with tax-deductible premiums and on-site medical clinics that provide more than de minimus medical care. The base does not include other fringe benefits, such as coverage under a separate policy for dental and/or vision care, or coverage for long-term care.

The threshold can be adjusted upwards for employers based on demographic characteristics. This adjustment would apply to employers with age and gender characteristics of employees significantly different from the national average. An upward adjustment can also be made for employees in risky professions defined in the statute, such as firefighters and paramedics, longshoremen and workers in construction, mining, agriculture, forestry and fisheries. An adjustment can also be made for retirees over 55 without Medicare coverage. For the latter two categories, retirees and high-risk professions, the dollar limits are increased by $1,650 for single coverage and $3,430 for family coverage.

The tax is legally imposed on the insurer in the case of group health plans offered by a fully insured employer. For contributions to a health savings plan (HSA) or an Archer MSA (medical savings account), the employer is responsible for the tax. For firms that are self-insured, the plan administrator is responsible. The expectation is that, in cases where firms retain their high-cost coverage, the tax would be passed on to employees through lower wages. In cases where employers reduce health insurance coverage to avoid the tax, wages would be expected to rise by the reduced benefit.

The Cadillac tax was enacted as part of the Patient Protection and Affordable Care Act of 2010, which made a number of changes in health insurance designed both to make affordable health insurance available to the uninsured and to contain the cost of health care. The Cadillac tax served two purposes: to raise revenue to finance costs of the Affordable Care Act (such as subsidies to low and moderate income families) and to discourage excessive spending on health care due to generous health insurance policies that limited the out-of-pocket costs (and thus health care prices) faced by consumers.

The Cadillac Tax and the Exclusion for Employer Sponsored Insurance (ESI)

The origins of the Cadillac tax are rooted in discussions to reform the tax subsidies for employer-provided health insurance. Amounts paid by firms on behalf of their employees for health insurance are excluded from wages and are subject to neither income nor payroll taxes. These health insurance benefits include purchase of group insurance on behalf of employees or self-insurance, where employers pay claims. Health coverage may also be selected as part of a “cafeteria” plan where employees choose among a menu of benefits. Deductible contributions may occur through specialized health savings accounts (HSAs) and flexible spending accounts (FSAs). All of these expenses can be excluded from taxable income.

These subsidies rank as one of the largest (if not the largest) income tax subsidies and, additionally, benefit from exclusions from payroll taxes (such as those paid for Social Security and Medicare). The value of these tax benefits was estimated by the Joint Committee on Taxation (2016) at $323.3 billion for 2016: $198.3 billion for exclusion from the income tax, $124.5 billion for exclusion from payroll taxes and $0.5 billion from exclusion from the additional Medicare tax of 0.9% on high income earners.

This exclusion has long been a part of the tax code. The Revenue Act of 1918 allowed the exclusion for employer provided health plans, and in 1943, the IRS ruled that direct contributions by employers to group health plans could be excluded from compensation. Miller (2014) links this ruling to a regulation by the War Labor Board that allowed employers to provide fringe benefits to avoid wage controls. The status of contributions to individual plans remained uncertain and IRS ruled in 1953 that they should be taxed. This ruling has a brief existence as Congress provided a broad exclusion of all employer contributions in 1954. Over time other types of benefits, such as flexible spending accounts and health savings accounts, were made exempt.

The exclusion has grown in relative size, compared to other tax expenditures and to the economy. When the first tax expenditure estimates for the income tax were compiled in 1968 (U.S Department of the Treasury 1969), the exclusion of medical insurance premiums, while a significant tax expenditure, was smaller than many other provisions: about a third of the size of the exclusion for pensions, and half the size of the deduction for charitable contributions. In the latest tax expenditure estimates (by the Joint Committee on Taxation 2015), it was the largest tax expenditure, 12% larger than pensions, and 3.3 times charitable contributions. This change in position was due to overall growth in the coverage of employer health insurance: the tax expenditure increased from 0.1% of GDP in 1968 to 1.1% in 2015. The tax expenditure for pensions also grew, although not as much, from 0.4% of GDP to 0.7%, while charitable contribution deductions remained about the same, at 0.25% of GDP.

Carasso (2005) plots the growth in the tax expenditure after the Tax Reform Act of 1986, through 2005, adjusted for both the consumer price index and the health cost price index, showing that a major reason for the growth of the exclusion is the rising cost of health care.

The exclusion of employer-sponsored insurances significantly reduces the relative price of health insurance. The magnitude depends on payroll taxes, which are 6.2% each on the employer and the employee for Social Security and 1.45% for Medicare (hospital insurance), as well as income taxes. While there is no cap on wages subject to Medicare taxes, wages for Social Security tax purposes are capped, currently (2016) at $118,500. Income tax rates are imposed at 10, 15, 25, 28, 33 and 39.6%. For certain high-income workers there is an additional 0.9% tax on wages. Based on the taxable income amounts, it is possible to have single individuals and two-earner married couples paying income tax at marginal rates as high as 28% while still being subject to the full payroll taxes. The most common marginal tax rate, however, is 15% (based on distributional data provided by the Joint Committee on Taxation 2009).

The assumption is generally made that, although half of the payroll tax is imposed by statute on the employer, the burden falls on workers through reduced wages. The lower wage causes the tax base to be smaller, so that the price differential cannot be calculated by adding the payroll and income tax rate. With fixed labour compensation (setting aside other fringe benefits), the wage can be determined by the relationship $W(1 + p) + B = F$, where W is the wage, p is the payroll tax rate, B is the health insurance benefit and F is a fixed amount. This relationship means that if benefits rise by one unit, wages fall by $1/(1 + p)$. When p is 0.0765 (the sum of the Social Security and Medicare tax rates for employees below the ceiling), wages fall by 7.1%, which is the amount of the payroll tax paid by the employer ($p/(1 + p)$). The employee also pays a tax of $p/(1 + p)$ and the income tax at $t/(1 + p)$, where t is the marginal tax rate. Thus the tax subsidy for health benefits relative to wages is $(2p + t)/(1 + p)$.

For employees below the Social Security payroll tax ceiling, p is 0.0765, and the subsidy, in

percentages, is 40.2, 37.4, 28.1, 23.5 and 14.2 for marginal income tax rates of 28, 25, 15, 10 and 0%. For employees above the ceiling, $p = 0.0145$, and the subsidy, in percentages, is 42.8, 38.2, 36.3 and 31.3, for tax rates of 39.6, 35, 33 and 28%. The first three of these also include the additional 0.9% tax on high incomes. Thus, for most employees the subsidies are close to or above 30%.

Data provided by the Urban Brookings Tax Policy Center (2016) indicated that 70% of the benefit of the tax exclusion accrues to the top 40% of the income distribution and 45% to the top quintile. These measures correspond (based on incomes in 2018 above $81,631 for the fourth quintile and above $143,318 for the highest quintile) to expected income tax rates of 25% or more. About 20% of the burden falls on the middle quintile, which would likely fall in the 15% or 25% income tax rate.

The payroll tax exclusion modifies the tendency of the tax subsidies to rise with income, because at the ceilings for Social Security, the payroll tax subsidy declines. At the same time, the subsidy arising from the exclusion from income of the payroll tax for Social Security is overstated, since the reduction in payroll taxes reduces future Social Security benefits. The Congressional Budget Office (2013) reported that individuals have lifetime Social Security benefits that are roughly equal to payments at a 3% discount rate. If subsidies below the payroll tax ceilings were restricted to the 1.45% Medicare tax (which is not tied to future benefits), the subsidy for tax rates from 28% down to 0% would be, in percentages, 30.5, 27.5, 17.6, 12.7 and 2.9%.

Subsidizing employer-provided health insurance has both desirable and undesirable features. Provision of health insurance suffers from an important market failure, adverse selection, where the inability of insurers to have full information on health status tends to overprice policies for healthier individuals and drive them out of the market. The pooling that arises in employer health plans help to reduce this market failure. In addition, in a private market, individuals with a preexisting health condition or health risks (such as age) may not be able to purchase affordable insurance, or any insurance at all, which may be undesirable from a social welfare perspective. At the same time, health insurance once purchased causes its own market failure, moral hazard, in that individuals who do not face the full cost of health care tend to overconsume it. In addition, because of health care coverage individuals may be less risk-averse.

Some participants in the Senate Committee on Finance roundtable discussion of health reform (2009) advocated eliminating or reducing the tax subsidies for employer-provided health insurance because of the effect on consumption and the "upside down" nature of the subsidy which tends to favour higher income taxpayers (especially if Social Security taxes are not considered).

Early on during the Senate Committee on Finance roundtable discussion (2009), Chairman Max Baucus made it clear that the elimination of health insurance subsidies was not on the table. At the same time, he expressed concern about its distributional effects and the incentive to buy too much health insurance. Members of this discussion group (committee members and outside experts) considered eliminating the exclusion for amounts above a dollar ceiling or a percentile of the actuarial value of the average plan.

Practically speaking, limiting the tax exclusion could face administrative challenges, including undesirable side effects for society. The pooling mechanism that was a benefit of employer health insurance creates problems for imputing income from employer-financed health insurance. If income were to be imputed based on the characteristics of the employees (the value of the insurance to them), the amount of income included could be onerous for some individuals, such as older employees. The simpler method of dividing costs by the number of employees means the imputed value subject to tax would depend on the health status of all the employees of the firm and could differ for employees with identical health plans and health conditions.

Whether for reasons of political optics or complications of imputing income, the Finance Committee ultimately proposed an excise tax on excess insurance costs to be paid by the insurance company (or the employer, in the case of self-insured plans). Because the tax is imposed at a flat rate, it

cannot offset the subsidy, which depends on graduated income taxes and flat payroll taxes with a ceiling on wages for Social Security.

The first proposal was at a 35 percent rate, but the rate was eventually raised to 40% and the tax was delayed until 2018. The original excise tax was not deductible for income tax purposes; revisions made in the Consolidated Appropriations Act of 2015 allowed deductibility and delayed the tax until 2020.

Penalties Imposed by the Cadillac Tax

The tax's success in discouraging high-cost health care plans with greater coverage depends on the size of the penalty. While a 40 percent rate appears to be high compared to the existing tax subsidies calculated earlier, it is not comparable to the existing subsidy rates because it is imposed on a tax-exclusive basis, in the same way as a sales tax (imposed on a base exclusive of the tax). Income and payroll tax subsidies are tax inclusive (imposed on a base inclusive of the tax). Thus a tax of 40% imposed on a tax inclusive basis is the equivalent of a 28.57 percent rate (0.4/1.4) on a tax-inclusive basis.

The excise tax is expected to be passed on to the worker as lower wages and will be offset by the tax savings on the lower wage. An additional dollar in benefits will cause wages to fall by 1.4/(1 + p), and this wage reduction will reduce the payroll and income taxes. The total tax effect reduction in wages is $-1.4(1-[(2p + t)/(1 + p)])$. This amount of reduced net wages can be decomposed into $1-(2p + t)/(1 + p)] + 0.4–0.4 [(2p + t)/(1 + p)]$. Tax effects include the tax subsidy on a dollar of benefits from the exclusion of benefits (noted above) of $(2p + t)/(1 + p)$ as well as a penalty from the excise tax and its accompanying wage offset of $0.4(1-[(2p + t)/(1 + p)])$. When $[(2p + t)/(1 + p)]$ is equal to 28.57%, the tax effects are zero, with the excise tax offsetting the subsidy from exclusion.

In many cases, the excise tax is too small to offset the initial subsidy, at least after the revision allowing deductibility of the tax. For the cases under the payroll ceilings where p is equal to 7.65%, the subsidy or penalty (denoted as a negative, showing a net tax), in percentages, is 16.3, 12.4, −0.6, −7.1 and −20.1 for marginal income tax rates of 28, 25, 15, 10 and 0%. For employees above the ceiling, $p = 0.0145$, the subsidy, in percentages, is 18.6, 12.3, 9.5 and 2.6, for tax rates of 39.6, 35, 33 and 28%. At a 15 percent rate, the income tax subsidy and the Cadillac tax penalty largely offset each other. For higher rates, the Cadillac tax is not large enough and a subsidy, albeit smaller, remains in effect. Within a payroll tax category, the initial subsidy from the tax savings for benefits is larger the higher the tax rate, while the Cadillac penalty is smaller, reflecting the offsetting effect of taxes on reduced wages on the tax.

Data provided by the Urban Brookings Tax Policy Center (2015) indicated that two-thirds of the burden of the Cadillac tax accrues to the top 40% of the income distribution and 37% to the top quintile. These measures correspond (based on incomes above $81,000 for the fourth quintile and above $143,000 for the highest quintile) to expected income tax rates of 25% or more. About 20% of the burden falls on the middle quintile, which would likely fall in the 15 or 25% income tax rate bracket. Thus, most high-cost plans would continue to receive a subsidy. The Cadillac tax burden is somewhat less concentrated in the higher income classes than the employer exclusion.

As noted earlier, the share of the payroll tax that finances Social Security benefits might be considered as funding an annuity, not a tax. In this case, the subsidy net of the payroll tax for marginal rates from 28% to 0% is, in percentages, 2.6, −1.5, −15.3, −22.2 and −36.9. Excluding this part of the payroll tax increases the regressiveness of the Cadillac tax (due to the more progressive income tax dominating income and payroll taxes) and results in a net penalty for most tax rates.

Allowing deductibility of the tax significantly reduced the size of the initial Cadillac tax. Wages and payroll taxes are deductible from the corporate or business profits tax base. If compensation is fixed at $W(1 + p)(1 - u) + B(1 - u) + .4B$, where u is the corporate tax rate, then the tax will cause a dollar of benefits to decrease wages by $1.4/[(1 - u)(1 + p)]$ without deductibility, rather than

0.4/(1 + p) with deductibility. The total effect can be decomposed into a tax of 0.4, a corporate tax on the fall in wages of 0.4u/[(1 − u)(1 + p)], an employer payroll tax savings which is deductible of p(1 − u)/[(1 − u)(1 + p)] and individual payroll and income taxes of (p + t)/[(1 − u)(1 + p)]. Setting u at 35%, for the cases under the payroll ceilings where p is equal to 7.65%, the subsidy or penalty (denoted as a negative, showing a net tax), in percentages, is 3.4, −1.1, −16.2, −23.6 and −38.7 for marginal income tax rates of 28, 25, 15, 10 and 0%. For employees above the ceiling, p = 0.0145, the subsidy, in percentages, is 5.7, −1.7, −4.9 and −12.3 for tax rates of 39.6, 35, 33 and 28%, with the first three including the 0.009 additional individual Medicare tax. The Cadillac tax, as initially designed, largely offset the subsidy from the exclusion of income and in many cases led to a net penalty.

Allowing deductibility provided more equal treatment of firms with differing tax rates (including tax exempt non-profit and government employers) but diminished the effect of the Cadillac tax in discouraging high-cost plans as well as reducing the revenue yield.

If the Social Security payroll tax is disregarded for this calculation, the subsidy rates for income tax rates from 28% to 0%, in percentages, are −12.3, −17.1, −33.0, −41.0 and −56.9 and thus all subject to penalties.

Growth of the Tax

The effect of the Cadillac tax depends on its burden and also on the scope of coverage. Because the Cadillac tax floor is indexed for inflation, the share of health insurance policies potentially subject to the tax will grow over time if the rate of growth in health insurance costs is greater. This expectation is evident from revenue effects as reported by the Congressional Budget Office (2016), with projected revenues rising from $2 billion in FY2020 to $20 billion in FY2025.

Lowry (2015) projects the growth in the percentage of plans covered using a lower growth assumption (4.6%), a moderate growth assumption (5.0%) and a high growth assumption (7.0%). Under the moderate growth scenario, in 2018, the tax would have affected about 10% of insurance plans for singles and about 8% for family plans; by 2025, these shares are over 20% for singles and around 20% for family plans and by 2030 the shares for single plans are around 40% and for family plans about 35%. Moreover, not only does the share of plans affected grow, but a greater share is subject to the tax over time.

The study also estimated the point at which the floor would be equal to the premium paid by the basic option Blue Cross Blue Shield (BCBS) plan offered by the Federal government (the lower cost of the two BCBS options). With moderate growth, for single plans, this point would be reached around 2030; with lower growth a year or two later, and with high growth by around 2023. For the family plan, the point would be reached around 2032 with moderate growth, around 2037 with low growth and around 2026 with high growth.

Geographic Differentials

One of the concerns with the Cadillac tax is the differential effects across states, since the tax threshold is not adjusted for geographical variations in health costs. Lowry (2015) also investigated these effects for 2018, using the lower growth scenario. For single plans, nationwide, 10.2% are subject to the tax. In Alaska, the share was projected at 29.6%, followed by Wyoming at 17.6%, Massachusetts at 15.8%, Montana at 15.7% and Delaware at 14.8%. Generally the states with larger shares were in the Northeast and Midwest, along with California. The smallest share was in Iowa at 4.7%, followed by Idaho, with 4.9% (although those data had a large standard error, suggesting caution), Hawaii and Alabama at 5.3%, Arkansas at 5.5% and Mississippi at 5.6%. For family plans with a national average of 6%, Alaska was 19.7%, followed by Connecticut at 12%, New Jersey at 9.7% and Maryland at 8.7%. States with higher shares were generally in the northeast. The states with the lowest shares were Idaho, at 1.6% (again with a large standard error),

Colorado at 2.3%, Louisiana at 2.4% and Mississippi and Arkansas at 2.5%. Generally, southern and western states (with the exception of Georgia, Nevada and Missouri) were below average.

Effect on Health Spending

One objective of the Cadillac tax was to address moral hazard: to discourage health plans that covered a larger proportion of health costs because consumers facing lower costs would be encouraged to consume too much. That effect would only occur if the tax caused employers to substitute wages for high cost plans.

Data from the Congressional Budget Office (2016), Gravelle (2015) and the Centers for Medicare and Medicaid Services (CMS) are used to derive a rough estimate of the potential effects on health spending. The Congressional Budget Office reports a $20 billion revenue gain for 2025, with 20–25% of that amount from the Cadillac excise tax and the remainder from changes in income and payroll taxes. Gravelle (2015) uses an assumption consistent with the Treasury Department that the sum of the individual payroll and income tax (p + t) for purposes of considering the Cadillac tax is 35%. If the payroll tax is 7.65%, it implies an income tax rate of 27.35% and the total subsidy amount (2p + t)/(1 + p) is 40%. If a 1.45% payroll tax is assumed, the value is 36%. Gravelle (2015) also uses an assumption that the out-of-pocket costs of health care are equal to 19% of the total.

The share of plans subject to the Cadillac tax should be smaller than the share of revenue because of the tax offset. If x is the share of plans that retain the high cost insurance, with others substituting the high cost insurance with wages, each dollar of revenue is equal to x0.4(1−(2p + t)/(1 + p))+(1 − x)(2p + t)/(1 + p). (These estimates assume that tax rates are the same for those plans that retain the high-cost insurance and those that eliminate it.) Using the mid-point of the JCT share, 0.225 and the mid-point of the subsidy rate, 0.38, 0.4x = 0.225 and 0.38*(1 − x) − 0.4x*0.38 = 0.775. Thus x is estimated at 0.199. Returning to the revenue amount and substituting the values for x and (2p + t)/(1 + p), each dollar of revenue is 0.199*0.4*(1 − 0.38) + 0.801*0.38. The share of revenues accounted for by the last term is 0.801/(0.199*0.4*(1 − 0.38) + 0.801*0.38), or 86%.

If the tax subsidy is 0.38, then each dollar of revenue represents 1/0.38, or $2.63 of income. The change in the subsidy for 2025 is 2.63*0.86*$20 billion, or $45 billion. This $45 billion is the change in the subsidy paid by insurance. Currently consumers face a price of P(1 − s), where P is the market price of health services and s is the share paid for by insurance (81%).

The Center for Medicare and Medicaid Services projects total health expenditures covered by private insurance for 2025 at $1752.6 billion. If the out-of-pocket share is 19%, the total expenditure is 1752.6/0.81 or 2163.7 and the out of pocket amount is $411.1 billion. So the percentage reduction in the subsidy is $45/$411 or 10.9%.

Consider two cases (as outlined in Gravelle 2015) of the effects of changing the subsidy of price and quantity in the health care market. Both rely on a demand elasticity of −0.2, with a supply elasticity of infinity in one case, and of 1.5 in the other. The percentage change in quantity is equal to 0.2 times the change in the market price plus the change in the subsidy (−10.9%). When the supply elasticity is infinity (i.e., perfectly elastic), the market price does not change and the percentage change in quantity is 2.2%. If the supply elasticity is 1.5 the market price falls with a reduction in quantity, and the change in the market price is the demand elasticity divided by the sum of the supply and demand elasticity times the percentage change in the subsidy. Thus, the market price falls by (0.2/(0.2 + 1.5)) times 10.9% or 1.3%. The quantity could be estimated from either the demand or supply curve, but using the supply curve, quantity falls by the supply elasticity times the percentage change in price, or 1.5 times (−1.3%), or 1.9%.

In the case of an infinitely elastic supply curve, total expenditure on health (in the private market) falls by 2.2%. In the case of a supply elasticity of 1.5, expenditure falls by the sum of the percentage

change in price (1.3%) and the percentage change in quantity (1.9%), or by 3.2%.

The effects of the Cadillac tax on health spending, absent revision, will tend to grow because the share of employer health insurance subject to the tax is projected to grow over time.

Conclusion

Although having the appearance of a high tax rate, the Cadillac tax rate, being expressed as a tax exclusive rate, allowed a net tax subsidy in most cases for retaining high-cost plans once it became deductible for income tax purposes. Eliminating the rule disallowing deductibility was a significant change. Nevertheless, economists at the Congressional Budget Office project that the tax will be effective at discouraging high-cost plans, perhaps because receiving income in cash is more desirable than minimizing premiums and co-pays.

According to the Department of the Treasury projections of effective tax rates, the income level of the recipient of the current benefits for these high cost plans is relatively high, although the employer-provided health insurance tax subsidy (or the Cadillac tax) does not rise with income as some other benefits (such as pensions and some itemized deductions) do. The tax should also reduce the size of the tax expenditure for employer-provided health insurance.

The reduction in high-cost health plans is expected, in turn, to reduce quantities, price, and expenditures on health costs for individuals covered by the private insurance market: the last could potentially fall by 3% or more. The Cadillac tax was perhaps the most important of the provisions of the Affordable Care Act targeted at reducing spending on health care, by discouraging use without raising market prices, and this aspect distinguishes it from other revenue-raising taxes and fees.

See Also

- ▶ Excise Taxes
- ▶ Health Insurance, Economics of
- ▶ Use of Experiments in Health Care

Bibliography

Carasso, Adam. 2005. Growth in the exclusion of employer health premiums. *Tax Notes*, June 27, p. 1697. At http://www.taxpolicycenter.org/sites/default/files/alfresco/publication-pdfs/1000794-Growth-in-the-Exclusion-of-Employer-Health-Premiums.PDF

Centers for Medicare and Medicaid Services, National Health and Expenditure Data. Visited Dec 9, 2016. https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/NationalHealthAccountsProjected.html

Congressional Budget Office. The 2013 long-term budget outlook, Sept 2013. https://www.cbo.gov/sites/default/files/113th-congress-2013-2014/reports/44521-LTBO-1Column_0.pdf

Congressional Budget Office. Private Health Insurance Premiums and Federal Policy. Feb 2016. https://www.cbo.gov/sites/default/files/114th-congress-2015-2016/reports/51130-Health_Insurance_Premiums.pdf

Gravelle, Jane G. 2015. *The excise tax on high-cost employer-sponsored health insurance: Estimated economic and market effects*, Congressional Research Service report R44159. Washington, DC: Library of Congress, Aug 20. https://fas.org/sgp/crs/misc/R44159.pdf

Joint Committee on Taxation. 2009. Background materials for Senate Committee on Finance roundtable on health care financing, JCX-27-09. At https://www.jct.gov/

Joint Committee on Taxation. 2015. Estimates of federal tax expenditures for fiscal years 2015–2019, JCX-141R-15, Dec 7. At https://www.jct.gov/

Joint Committee on Taxation. 2016. Exclusion for employer-provided health benefits and other health-related provisions of the internal revenue code: Present law and estimates, JCX-25-16, Apr 12. At https://www.jct.gov/

Lowry, Sean. 2015. *The excise tax on high-cost employer-sponsored health coverage: Background and economic analysis*, Congressional Research Service report R44160. Washington, DC: Library of Congress, Aug 20. https://fas.org/sgp/crs/misc/R44160.pdf

Miller, Tom. 2014. Kill the tax exclusion for health insurance. *National Review*, Aug 19. At http://www.nationalreview.com/article/385704/

U.S. Congress, Senate Committee on Finance. Roundtable discussion on expanding health care coverage, 111th Cong., 1st Sess., May 5, 2009. At http://www.finance.senate.gov/download/roundtable-discussions-on-comprehensive-health-care-reform

U.S. Department of the Treasury. Annual report of the Secretary of the Treasury on the state of the finances for the fiscal year ended June 30, 1968 (1st Congress, 1st Session, House Document No. 3, 1969).

Urban-Brookings Tax Policy Center. Table T15–0096 – Repeal Cadillac tax, premiums at post-Cadillac tax levels, distribution of federal tax change by expanded cash income percentile, 2018, July 23, 2015. http://www.taxpolicycenter.org/model-estimates/repealing-excise-tax-high-cost-health-plans/repeal-cadillac-tax-premiums-post

Urban-Brookings Tax Policy Center. T16–0159 – Tax benefit of the exclusion of employer-sponsored health benefits and deduction for self-employed health insurance premiums by expanded cash income percentile, 2016. http://www.taxpolicycenter.org/model-estimates/individual-income-tax-expenditures-july-2016/t16-0159-tax-benefit-exclusion-employer

Cairnes, John Elliott (1823–1875)

R. D. Collison Black

Keywords

Cairnes, J. E.; Deductive method; Fawcett, H.; Jevons, W. S.; Land tenure; Malthus's theory of population; Mill, J. S.; Rent control; Slavery; Subjective theory of value; Verificationist methodology; Wages fund

JEL Classifications

B31

Cairnes was born at Castlebellingham, County Louth, Ireland. At the height of his career he was probably the best-known political economist in England after John Stuart Mill, whose friend and associate he was from 1859 onwards; but his interest in economic questions developed relatively late, after periods spent working in his family's brewing business and in journalism. In 1856 he competed in the examination by which the Whately professorship of political economy at Trinity College, Dublin, was then filled, and was appointed for a five-year term. In 1859 he was also appointed Professor of Political Economy and Jurisprudence at Queen's College, Galway, a post which he held until 1870. However, he employed a deputy to perform his duties in Galway after he himself moved to London in 1865. In 1866 he became Professor of Political Economy at University College, London, but was forced to resign in 1872 by the progress of the rheumatic disease which left him almost completely paralysed before his death in 1875.

Cairnes has often been described as 'the last of the classical economists'. He always worked within the framework of the Ricardo–Mill tradition, devoting himself to refining and strengthening it and seeing no necessity for any radical reform or reconstruction. Within these self-imposed limits and in a career of less than 20 years as a professional economist, he succeeded in making contributions to both theoretical and applied economics which earned him a high reputation among his contemporaries and a definite place in the history of economic thought.

Cairnes's first work in economics proved to be one of his most enduring contributions to the subject. This was *The Character and Logical Method of Political Economy* (1857; 2nd edition, 1875) which is still regarded as one of the best statements of the verificationist methodology of the English classical school. Following the lines laid down by Senior and Mill, Cairnes stressed the neutrality of economic science, emphasized the value of the deductive method and characterized the subject as a hypothetical science 'asserting, not what *will* take place, but what *would* or what tends to take place' ([1857] 1875, p. 55).

It was in the use of the deductive method to develop the central areas of economic theory that Cairnes's main interest came to lie. Yet it was through his work on applied economics and current issues of policy that he first came to be nationally and internationally known. In September 1859 Cairnes published the first of a series of 'Essays towards a solution of the Gold Question' in which he sought to 'apply the principles of economic science' in an attempt to 'forecast the directions in which the course [of trade and prices] would be modified by the increased supplies of gold'. This a priori approach was almost precisely the opposite of that used by Jevons to deal with the same problem, but their results coincided remarkably.

It was another application of this approach which first made Cairnes's work known to a much wider audience. In *The Slave Power* (1862) he sought to explain on economic grounds the appearance of slavery in the southern parts of the United States, tracing out both the conditions for and the consequences of the operation of a slave economy. As an indictment of the political economy of the

Confederate States it strongly influenced public opinion in Britain towards support of the Northern states in the American Civil War.

Between 1864 and 1870 Cairnes wrote a number of articles on the problems of land tenure in Ireland, in which he argued in favour of proposals to fix rent by law and contended that this was not inconsistent with classical rent theory. There is evidence that his views on this and other questions of the day, such as Irish university education, exerted considerable influence on (and through) Mill and Fawcett.

Cairnes's most important contribution to economic analysis, *Some Leading Principles of Political Economy Newly Expounded* (1874), was also to be his last work and that by which he came to be most widely known and judged. In it he restated, but with significant modifications, the essentials of classical doctrine on the central questions of value, distribution and international trade. His most important innovation was to show that the existence of 'non-competing groups' in labour markets implied that the cost of production theory must be supplemented by the analysis of reciprocal demand in the theory of domestic as well as international values.

Nevertheless his unsympathetic review of Jevons's *Theory of Political Economy (Fortnightly Review,* N.S., vol. 11, 1872) showed that he lacked interest in and understanding of the subjective approach to value theory which was then developing. Cairnes's treatment of distribution in the *Leading Principles* echoed Mill in showing sympathy for the position of the labourer combined with pessimism based on acceptance of Malthusian population theory; but it was chiefly notable for an elaborate but ultimately unsuccessful attempt to rehabilitate the wages-fund doctrine abandoned by Mill himself in 1869. The verdict of Schumpeter (1954, p. 533) still seems appropriate: Cairnes 'expounded the old analytical economics and explicitly distanced himself from the new'.

Selected Works

1857. *The character and logical method of political economy*. London: Macmillan; 2nd ed., 1875; repr. 1888.

1862. *The slave power: Its character, career, and probable designs: Being an attempt to explain the real issues involved in the American contest*. London: Macmillan; 2nd ed., 1863.

1873. *Political essays*. London: Macmillan.

1873. *Essays in political economy, theoretical and applied*. London: Macmillan.

1874. *Some leading principles of political economy newly expounded*. London: Macmillan.

Bibliography

Black, R.D. Collison. 1960. Jevons and Cairnes. *Economica* 27: 214–232.

Boylan, T.A., and T.P. Foley. 1984. John Elliott Cairnes, John Stuart Mill and Ireland: Some problems for political economy. In *Economists and the Irish economy*, ed. A.E. Murphy. Dublin: Irish Academic Press.

O'Brien, G. 1943. J.S. Mill and J.E. Cairnes. *Economica* 10: 273–285.

Schumpeter, J.A. 1954. *History of economic analysis*. New York: Oxford University Press.

Weinberg, A. 1970. *John Elliott Cairnes and the American Civil War*. London: Kingswood Press.

Calculus of Variations

Morton I. Kamien

JEL Classifications
C61

The development of the calculus of variations is attributed to Euler and Lagrange, although some of it can be traced back to the Bernoullis. A history of the calculus of variations is provided by Goldstine (1980). The calculus of variations deals with the problem of determining a function that optimizes some criterion that is usually expressed as an integral. This problem is analogous to the differential calculus problem of finding a point at which a function is optimized, except that the point in the calculus of variations is a function rather than a number. The function over which the optimum is sought is usually

restricted to the class of continuous and at least piecewise differentiable functions.

A typical calculus of variations problem is of the form

$$\max_{x(t)} \int_{t_0}^{t_1} F[t, x(t), x'(t)]\mathrm{d}t. \ \text{s.t.} x(t_0) = x_0, \quad (1)$$

where $x'(t) = \mathrm{d}x/\mathrm{d}t$, and t, $x(t)$, and $x'(t)$ are regarded as independent arguments of the function F. The necessary conditions for $x^*(t)$ to maximize (1) are the Euler equation

$$F_x = \mathrm{d}F_{x'}/\mathrm{d}t, \quad (2)$$

F the Legendre condition

$$F_{x'x'} \leq 0 \quad (3)$$

and the transversality conditions

$$F_{x'} = 0 \text{at} t_1, \ \text{if } x(t_1) \text{ is free}, \quad (4a)$$

$$F - x'F_{x'} = 0 \text{ at } t_1, \ \text{if } t_1 \text{ is free}, \quad (4b)$$

where F_x and $F_{x'}$ refer to the partial derivatives of F with respect to x and x', respectively, and $F_{x'x'}$ is the second partial derivative of F with respect to x and x'. The Euler Eq. 2 is in general a nonlinear second order differential equation. The initial condition $x(t_0) = x_0$ and the transversality condition (4a) provide the means for determining the two constants of integration that arise in solving the Euler equation. The optimal value of the upper limit of integration, t_1, if it can be chosen, is determined by the transversality condition (4). The problem posed in (1) can be extended to include additional arguments of the function F, to include a variety of additional constraints, and to involve double integrals (see Kamien and Schwartz 1981). Concavity of F with respect to $x(t)$ and $x'(t)$ assures that the necessary conditions are also sufficient.

The earliest application of the calculus of variations to the analysis of an economic problem appears to have been attempted by Edgeworth (1881), who seems to have been greatly impressed by its successful employment in deriving some of the basic laws of physics. He sought to employ it to find a function for distributing income and assigning work among the members of society so as to maximize total social welfare. Many applications of the calculus of variations to economic problems have been conducted since then, a few of which will be described.

As the calculus of variations deals with the problem of finding a function or a path that maximizes some criterion, its major application in economics has been to problems involving optimal decision making through time where an entire course of actions is sought rather than a single action. One of the earliest and most influential applications along these lines is by Ramsey (1928). The question he addressed is how much should a nation save out of its national income through time so as to maximize its overall welfare over time. Ramsey argued that the discounting of future utilities was 'ethically indefensible' as it means that we give less weight to the utility of future generations than to our own. He posited, therefore, a maximum level of net utility, the utility of consumption minus the disutility of work, that he called bliss. This bliss level of utility is the asymptotic limit of the achievable level of net utility. Ramsey then sought the savings rate through time that would minimize the integral over the indefinite future of the difference between the bliss level of utility and the actual net utility level at each point in time, subject to the constraint that savings plus consumption equal total output at each instant of time. The rule he derived for the optimal savings rate, through the Euler equations, is that the 'rate of saving multiplied by the marginal utility of consumption should always equal bliss minus actual rate of utility enjoyed'. This is essentially a marginal sacrifice today equals marginal benefit tomorrow rule. The rationale for taking the upper limit of integration to be infinite in the objective function is that while individuals have finite lives, society as a whole goes on forever. Ramsey also took up the case where future utilities are discounted at a constant positive rate and derived what may be regarded as the fundamental equation of optimal consumption through time, namely that the proportionate rate of change of marginal utility of consumption should equal the

C

difference between the marginal productivity of capital and the rate at which future utility is discounted. The Ramsey model became the basis for optimal growth theory that was intensely investigated in the late 1950s and 1960s.

Strotz (1956) addressed the question of the circumstances under which an individual would continue today to follow the optimal consumption plan through time that he had determined at an earlier date. In other words, he asked for the conditions under which an optimal consumption plan through time would be consistent. He found the necessary and sufficient conditions for consistency to be that 'the logarithmic rate of change in the discount function must be constant'. Exponential discounting at a constant rate satisfies this criterion.

Yaari (1965) addressed the question of an individual's optimal consumption plan through time when his lifetime is uncertain. He also allowed for the possibility that the individual derives utility from a bequest to his heirs. Yaari found that a major effect of the presence of uncertainty about one's lifetime is the same as an increase in the rate at which future utilities are discounted. Thus, the 'effective' rate at which future utilities are discounted has a risk premium term added to the discount rate in the absence of uncertainty about one's lifetime. The risk premium term is the instantaneous conditional probability of dying in the next instant given survival to the present. The presence of the risk premium means that the rate of consumption at any point in time is higher than it is in its absence. Uncertainty about one's lifetime increases one's rate of current spending, if there is no bequest motive.

While Ramsey applied the calculus of variations to the problem of optimal savings through time, Evans (1924) appears to have been the first to have employed it for determining the optimal rate of output through time. Evans used, as his vehicle for making the problem of choosing the level of output so as to maximize a monopolist's profit over an interval of time nontrivial, i.e. just simple maximization of profit at each instant of time, the assumption that the demand function for a good depended both on its current price and the rate of change of price. In particular, he assumed that the demand function was linear in price and its first derivative, and that the cost of production was a quadratic function of the level of output. Under these assumptions Evans sought the level of production that would maximize the integral of profits over a finite horizon. He was able to characterize this path and to show that a particular solution to the second order differential equation stemming from the Euler equation was the static monopoly profit maximizing level of output. Indeed, it is not difficult to show that when the problem is posed as one of maximizing the present value of an infinite horizon profit stream that the static monopoly profit maximizing level of output and the corresponding monopoly price constitute a steady-state towards which the output and price paths converge through time. This, of course, is intuitively plausible, as in the steady-state the rate of change of price with respect to time is zero, and so the demand function depends only on the current price level. Evans's work was extended by Roos (1925) to the case of duopolistic producers of a homogeneous product seeking to maximize their individual profits through time. The Roos paper may be regarded as the earliest analysis of what has come to be known as a differential game (see Fershtman and Kamien 1987).

The last paper that deserves special mention because of its important application of the calculus of variations is Hotelling's (1931), dealing with the rate at which a mineral resource such as coal, copper or oil should be extracted from a mine and sold so as to maximize the present value of its profits. Hotelling derived the fundamental equation for optimal extraction, under competitive production of the resource, namely that the extraction rate be such as to equate the percent change in price through time with the rate of interest at each instant in time. The intuitive reason for this is that if the percent change in the price of the resource exceeds the interest rate then it pays to extract and sell more today, because the alternative of extracting less and earning the interest on the revenue from that level of extraction yields less. The increase in the current rate of extraction, however, causes price to decline until the percent change in the price through time is equalized with the rate of interest. A similar analysis yields that current extraction will decline if

the percent change in price is below the interest rate, which in turn will cause price to rise until equality is achieved. Along the optimal extraction path the mine owner is just indifferent between extracting an extra unit of resource today and extracting it tomorrow. A similar analysis can be carried out for a monopolistic mine owner, with the percent change in marginal revenue through time being equated with the interest rate.

There have been a very large number of applications of the calculus of variations since these early ones. Many have employed optimal control methods and dynamic programming methods, both of which constitute generalizations of the calculus of variations. As long as decision making though time is regarded as an important subject of economic analysis, the calculus of variations will continue to find use in economics.

See Also

- ▶ Edgeworth, Francis Ysidro (1845–1926)
- ▶ Evans, Griffith Conrad (1887–1973)
- ▶ Ramsey model
- ▶ Roos, Charles Frederick (1901–1958)

Bibliography

Edgeworth, F.Y. 1881. Mathematical psychics. Reprinted, New York: Augustus M. Kelley, 1967.

Evans, G.C. 1924. The dynamics of monopoly. *American Mathematical Monthly* 31: 75–83.

Fershtman, C., and M. Kamien. 1987. Dynamic duopolistic competition with sticky prices. *Econometrica* 55: 1151–1164.

Goldstine, H.H. 1980. *A history of the calculus of variations*. New York: Springer.

Hotelling, H. 1931. The economics of exhaustible resources. *Journal of Political Economy* 39: 137–175.

Kamien, M.I., and N.L. Schwartz. 1981. *Dynamic optimization*. New York: North-Holland.

Ramsey, F.P. 1928. A mathematical theory of saving. *Economic Journal* 38: 543–559.

Roos, C.F. 1925. A mathematical theory of competition. *American Journal of Mathematics* 47: 163–175.

Strotz, R.H. 1956. Myopia and inconsistency in dynamic utility maximization. *Review of Economic Studies* 23: 165–180.

Yaari, M.E. 1965. Uncertain lifetime, life insurance and the theory of the consumer. *Review of Economic Studies* 32: 137–150.

Calibration

Edward C. Prescott and Graham V. Candler

C

Abstract

The methodologies used in aerospace engineering and macroeconomics to make quantitative predictions are remarkably similar now that macroeconomics has developed into a hard science. Theory provides engineers with the equations, with many constants that are not well measured. Theory provides macroeconomists with the structure of preference and technology and many parameters that are not well measured. The procedures that are used to select the parameters of the agreed upon structures are what have come to be called 'calibration' in macroeconomics.

Keywords

Calibration; Elasticity of intertemporal substitution; Equity premium; Impatience; Lucas critique; Measurement; Neoclassical growth theory; Risk aversion; Total factor productivity

JEL Classifications

D4; D10

What is calibration? In the dictionary definition, calibration is the act of calibrating a measurement instrument so that it gives the correct measurement for some known conditions. When calibrating a thermometer that will be used to measure the air temperature, calibration would involve setting it to read 100 degrees Celsius when submerged in boiling water at sea level and zero degrees when submerged in ice water. Because the boiling point of water varies with altitude, the calibration would be different in Mexico City, which is more than a mile above sea level.

Sometimes macroeconomists calibrate a measurement instrument – that is, a model – in this narrow sense. But calibration has gained a broader meaning in economics and is what

macroeconomists do when using theory to derive *quantitative theoretical inference*. Prescott emphasizes that calibration is not estimation. Calibration is a process that uses theory to construct a model – that is, an instrument – which will be used to provide a quantitative answer to a question.

Clearly, instruments are not measured; rather, they are calibrated so that they can be used to accurately answer quantitative questions. The nature of questions varies. Examples of questions are as follows: what is the welfare benefit or cost of changing the currently employed policy arrangement to another one? What will happen to a spacecraft when it enters the atmosphere of Mars?

To predict the quantitative consequences of a particular policy, theory and observations are used to select a model economy, and the equilibrium behaviour of that economy is determined for the proposed policy. Theory provides a set of instructions for selecting the model economy. *This selection process is what calibration in economics has come to mean*. Needless to say, the nature of the application of theory and the availability of economic statistics dictate which model economy is selected.

Before proceeding, a little history of the development of macroeconomics is needed. The modern national accounts were developed by the NBER staff in the 1920s, with Simon Kuznets playing the leading role. In the 1950s and 1960s, macroeconomists searched for *the dynamic system* governing the behaviour of these accounts. The controls for this dynamic system were policy actions. Not having much theory, this activity was largely empirical. Macroeconomists would write down a parametric set of models and find the one that best *fitted* the national accounts, augmented with other statistics. This search for *the dynamic system* failed because, as established in the Lucas critique, the existence of such a policy invariant dynamic system is inconsistent with dynamic economic theory.

The failure of this search led to a vacuum in quantitative macroeconomics. The profession did not want to go back to conjecturing and storytelling that characterized pre-war business cycle theory. As a result, the 1970s was a frustrating decade for quantitative macroeconomists given the failure of the empirical approach and the lack of needed tools and theory to quantitatively study macroeconomic behaviour.

This vacuum was filled in the early 1980s when the extended neoclassical growth model was used to study business cycles. The national accounts had to be modified to be consistent with the model. The most important modification in the study of business cycles is treating consumer durable expenditures as an investment and imputing consumption services to the stock of consumer durables as is done for owner-occupied housing. The secular growth observations with constancy in shares of output led to a constant elasticity structure with share and elasticity parameters. The fact that capital share of income displayed no trend even though the relative price of labour increased secularly led to a unit elasticity of substitution aggregate production function with share parameters equal to income shares. The depreciation rate, for example, was calibrated to average depreciation share of product. The national accounts use prices of used capital goods to estimate depreciation.

This methodology is used in virtually all quantitative theoretical aggregate studies. We emphasize that quantitative theoretical research and empirical research are fundamentally different activities and fundamentally different tools are needed. If the objective of the research is to derive the quantitative implications of the neoclassical growth theory for business cycle fluctuations, the use of statistical tools to select the parameters that best fit the business cycle observations is not sound scientific practice.

In this short article macroeconomist Prescott will describe what he does when addressing macroeconomic issues and aerospace engineer Candler will describe what he does when addressing the problem of making predictions of what will happen when a capsule enters the atmosphere of Mars. These predictions are relevant to the design of the capsule. Prescott will conclude by comparing the approaches and argue that these scientific approaches are essentially the same. We begin with what aerospace engineers do so that comparison can be made with what they do and what macroeconomists do.

Candler: The Aerospace Engineer

I work in the field of aerodynamics, and specifically I try to predict what happens when a spacecraft enters the atmosphere of a planet. For example, one of my current projects involves predicting how the Mars Science Laboratory capsule will fly as it enters the Martian atmosphere. What is the peak heat transfer rate to the spacecraft? How much heat shield is required to protect it from the extremely high temperature gas that surrounds it during atmospheric entry? Will it produce enough lift so that it will fly along the planned trajectory? Will the uncertain state of the atmosphere cause the capsule to veer off course? These questions must be answered to a known level of accuracy before the spacecraft can be designed. Failure to predict heating levels or aerodynamic performance can result in a well-publicized and expensive loss of the mission. At the same time, excessive conservatism in the design reduces the useful payload of the spacecraft and increases the cost of the mission.

How do we go about modelling this complex problem? We cannot fly a statistical ensemble of missions and empirically extrapolate to the flight conditions of interest. Instead, we must rely on ground-based wind-tunnel testing and theory-based simulations. However, experiments have a number of limitations: it is impossible to test the full-scale capsule; it is usually impossible to produce the actual flight conditions; and we cannot produce the actual intense heating levels for realistic periods of time. On the other hand, we can use numerical simulations to predict the flow field around the full-scale spacecraft at critical points in the entry trajectory. In principle, these calculations can predict the heat transfer rates and aerodynamic forces, and provide accurate data for the spacecraft designers. Of course, these simulations are only as accurate as the underlying equations being solved, and herein lies the problem. We cannot rely on purely empirical measurements to test a spacecraft design, yet simulations require a set of governing equations that must be validated by realistic flight experiments.

Interestingly, the basic set of governing equations that describes the flow over a spacecraft entering a planetary atmosphere is well established. However, there are many parameters in these equations that are the subject of intense debate within my field. We do not have an accurate understanding of the chemical reaction rates in the flow field; we do not know how to model transition to turbulence in the flow near the surface; we cannot predict how much turbulent flow enhances the heat transfer rate; and we do not understand how the high-temperature gas interacts with the spacecraft surface. A complete model of the flow over a spacecraft entering the atmosphere of Mars has well over 100 model constants that must be determined before the equations are fully specified. Clearly, with our limited experience base and with the limitations of the ground-based testing facilities, it is fundamentally impossible to determine these model constants with the available data. Rather, we must impose a rigorous theoretical basis for the choice of these model parameters. Also, we must understand the sensitivity of the critical results (heat transfer rate and aerodynamic forces) to the choice of the parameters. For example, there is no sense in investing a lot of time and money to accurately determine a model parameter that has a one per cent effect on the lift at relevant conditions.

So what do we do? We attack the problem from two sides. First, we break the full problem into well-defined parts and use theory and experiment to determine specific parameters under controlled conditions. For example, we might be concerned with how high-temperature oxygen molecules attack a particular heat-shield material. We would commission experiments to address this specific issue at conditions that are as close as possible to the flight conditions. Typically, it is impossible to exactly reproduce the conditions, and we would then perform experiments in different test facilities to help bound the parameters. Theory would then be used to extrapolate from the test conditions to those encountered in flight. We always try to use a theoretical basis to provide discipline to this process. We never perform atheoretic variations of parameters to try to match the data – if it is necessary to break the laws of physics, there is usually something wrong!

The second approach to modelling the flow field is to determine what parameters really matter to the design. A very useful approach is to use theory and experience to bound the range of all parameters in the model. Then a large number of simulations are performed, sampling from the distribution of each parameter. With enough simulations, it is possible to determine the sensitivity of the spacecraft design to each of the modelling parameters. Usually with this parametric uncertainty analysis it is possible to isolate several critical parameters that require particular attention. For example, Wright, Bose and Chen (2007) determined that eight modelling parameters out of several hundred were responsible for 90% of the uncertainty in the design of a proposed spacecraft. New experiments were then designed and carried out to reduce the uncertainty in these critical parameters.

Another engineering perspective is worth noting. We fully recognize that our representation of the world will never be 100% per cent accurate. Rather, we must quantify the level of accuracy of a given model and determine if we can fly a mission with that implied level of risk. We must quantify levels of uncertainty in a design and recognize that a spacecraft that will never fail will be excessively expensive or will carry so little payload as to be worthless. Thus, there is a calculated risk associated with the uncertainty in our modelling parameters. Of course, we try to reduce this uncertainty, but ultimately we are always forced to live with some level of risk if we want to fly an interesting mission.

Prescott: The Macroeconomist

The selection of parameters in quantitative theory is not measurement. However, quantitative theory is often useful in measurement. It is also useful in making predictions and in accounting for observations. Some examples of successful application are as follows.

The Lucas (1978) asset pricing model with the Markov process on the growth rate of endowments places restrictions on the joint behaviour of asset returns and consumption given two parameters that specify the stand-in household's preference ordering. The first parameter is the degree of risk aversion and the second parameter is the degree of impatience. These restrictions hold in worlds in which there are no transaction costs, no taxes, and no intermediation costs. Whether abstracting from certain factors is reasonable or not depends upon the question.

Mehra and Prescott (1985) used this asset-pricing model economy to estimate how much of the historical equity premium is a premium for bearing aggregate risk. We selected a Markov aggregate endowment growth-rate process whose first two moments matched the historical experience. We used observations and theory to restrict the values of the two preference parameters, including numerous observations on household behaviour. This process of restricting these parameters is part of the calibration process. We found that only a small part of the historical equity premium was a premium for bearing aggregate risk for *any* value of the parameters in the restricted range. This model economy is ill suited for measuring the curvature and impatience parameter of the stand-in household, but it was well suited for determining how much of the historical equity premium is for bearing aggregate risk.

I turn now to a case where a key economic parameter was estimated accurately using a calibrated *set* of model economies. The neoclassical growth model used to study business cycles was used to estimate the leisure intertemporal elasticity of substitution parameter. This parameter is crucial for evaluating tax policies. Because the income and substitution effects roughly offset secularly, balanced growth observations say nothing about the magnitude of this elasticity parameter. If the neoclassical growth model is accepted as a good abstraction for studying business cycles, business cycle observations tie down this parameter. But the profession was reluctant to accept this theory as a useful one for studying business cycles and therefore did not accept the business cycle-based estimate of this elasticity.

This important parameter was tied down by cross-country and cross-time observations on tax rates and labour supply. Tax rates, broadly defined to be those features of policy that affect the households' budget constraint, account for virtually all

the large differences in labour supply across the large advanced industrial countries and across time for France, Italy and Germany. That this estimate is the same one found in the study of business cycles gave confidence to the view that business cycles are in major part optimal responses to real shocks including productivity, taxes, and terms of trade. As established theory and measurement were used in this study, this is calibration.

I turn now to a specific application of the neoclassical growth model to the study of the aggregate value of the stock market, which also entailed calibration. The study that began in late 1999 was motivated by the question of whether the stock market was overvalued and about to crash. At that time people did not know how to use this theory to obtain an accurate answer to this question and relied on historical relations such as price–earnings ratios to answer the question.

To address this issue neoclassical growth theory as developed in the study of business cycles was used. The model economy had to be modified in three important ways. First, there had to be at least two production sectors, a corporate and a non-corporate sector. To have a reason for having two producing sectors, the outputs of the sectors must be different and must be aggregated in some way. McGrattan and Prescott (2005) use the standard procedure of introducing an aggregator of the sector outputs that produces a composite final output good. This aggregator has a share parameter that must be *calibrated* to some observation. The observation selected is the average relative outputs of these two sectors. This is a crucial dimension for the model to mimic reality, given the issue being addressed. The conclusion turned out to be insensitive to the elasticity of substitution between these inputs, which was fortunate given there is not good information on this elasticity. Second, the tax and regulatory system had to be modelled explicitly. For example, we set the model's tax rate on corporate distributions equal to the average marginal tax rates on distributions. This is calibration because in the model world this tax rate is the same for all individuals when in fact it is not. Third, we deal with the fact that corporations have large stocks of unmeasured productive assets and that these assets are an important part of the value of corporations, being stocks of knowledge resulting from investment in research and development, organization capital and brand capital. We figure out how to estimate this stock of unmeasured capital using national account data and the equilibrium conditions that the after-tax return on measured and unmeasured capital are equal.

A theory is tested through successful use. The theory correctly predicts the great variation in the value of the stock market in relation to GDP, which varied by a factor of 2.5 in the United States and by a factor of three in the United Kingdom in the 1960–2000 period. Little of this variation is accounted for by the obvious factors, namely after-tax earnings in relation to GDP and the debt–equity ratio, which varied little over time. The secular behaviour of the stock market value, with its large variation in relation to gross national income, turned out to be as predicted by theory and is not due to animal spirits.

Another example of successful calibration is Hayashi and Prescott (2002), who examined why Japan lost a decade of growth. The neoclassical growth model used in their study is the one used in the study of business cycles. The exogenous parameter paths were working-age populations, capital income tax rates, and total factor productivity parameters (TFP). The TFP parameters were determined residually from the production function given the quantities of the factor inputs and the output. Given these exogenous elements the equilibrium path was computed. The finding is that the Japanese economy behaved as predicted by the theory. The reason for the lost decade of growth was the failure of TFP to grow. This led to the important question of why Japanese TFP failed to grow as it did in western Europe and North America in this period.

Similarities and Differences Between Aerospace Engineering and Macroeconomics

Both Candler and Prescott study and model aggregate phenomena. Neither can find the answers empirically through trial and error and both must rely on theoretical computer simulations restricted

by measurement. We both test for the robustness of our predictions when making predictions as to what will happen in situations never experienced. In one case the prediction is what will happen to a spacecraft that will be sent to Mars. In the other case it is what will be the consequences of implementing a proposed policy arrangement. Both rely on established theory and measurement to draw quantitative inference.

A difference is that the engineers have the equations, while macroeconomists have statements about preferences and technology. A consequence of this is that macroeconomists have the added step of determining the equilibrium equations of their model. Another minor difference is that computational intensity is much greater in aerospace engineering than in macroeconomics.

See Also

- ▶ Financial Market Anomalies
- ▶ Kydland, Finn Erling (1943–)
- ▶ Lucas Critique
- ▶ Real Business Cycles
- ▶ Recursive Competitive Equilibrium

Acknowledgment We thank Gary Hansen, Ellen McGrattan, Berthold Herrendorf, Lee Ohanian, and Bob Lucas for comments. We are responsible for all views expressed.

Bibliography

Hansen, G.D. 1985. Indivisibility and the business cycle. *Journal of Monetary Economics* 16: 309–327.

Hayashi, F., and E.C. Prescott. 2002. The 1990s in Japan: A lost decade. *Review of Economic Dynamics* 5: 206–235.

Kydland, F.E., and E.C. Prescott. 1982. Time to build and aggregate fluctuations. *Econometrica* 50: 1345–1370.

Lucas, R.E. Jr. 1978. Asset prices in an exchange economy. *Econometrica* 46: 1429–1445.

McGrattan, E.R., and E.C. Prescott. 2005. Taxes, regulations, and the value of U.S. and U.K. Corporations. *Review of Economic Studies* 72: 767–796.

Mehra, R., and E.C. Prescott. 1985. The equity premium: A puzzle. *Journal of Monetary Economics* 15: 145–161.

Wright, M.J., D. Bose, and Y.K. Chen. 2007. Probabilistic modeling of aerothermal and thermal protection material response uncertainties. *AIAA Journal* 45: 399–410.

Cameralism

H. C. Recktenwald

Keywords

Becher, J. J.; Cameralism; Fiscal jurisprudence; Justi, J. H. G. Von; Mercantilism; Sonnenfels, J. Von; Utilitarianism

JEL Classifications

B1

Cameralism is the specific version of mercantilism taught and practised in the German principalities (*Kleinstaaten*) in the 17th and 18th centuries. Becher (1635–82), von Justi (1717–71) and von Sonnenfels (1732–1817) are the principal figures who contributed to a vast cameralist literature of about 14,000 titles (Humpert, 1935). The subject matter of *Kameralismus* reflected the political and economic phenomena and problems in the German territorial states. As a branch of 'science' it is a fiscal *Kunstlehre,* that is, the practical art of how to govern an autonomous territory efficiently and justly via financial measures designed to fill the state's treasury. Its subject matter includes economic policy, legislation, administration and public finance. While there is no unifying analytical foundation of cameralism, it did develop in two distinct phases (a younger and an older branch) with varied emphasis on its different elements, and since the rising state was, in theory and reality, the focus and *ultima ratio* of political, economic and ethical (occasionally promotive) speculation, cameralism takes on a unitary form (*Gestalt*) only when viewed in retrospect.

The term 'cameralism' itself originates in the management of the state's or prince's treasure (*Kammer, caisse, camera principis*), seen as the principal instrument of economic and political power. In the age of enlightened absolutism, German–Austrian cameralism, based on a somewhat obscure natural-law philosophy, emphasized the paternalistic character of the governments'

centralized fiscal policy (not, as is sometimes mistakenly thought, a Keynesian short-run instrument but rather a regulator for development which was to serve the general happiness of the subjects (*Untertanen*), that is, an eudaemonistic utilitarianism). English and French mercantilism, on the other hand, stressed much more the wealth or 'riches' of the sovereign as an end.

The princely bureaucrats had been trained in their own universities (for example, Halle, Frankfurt/Oder, Vienna) in 'fiscal jurisprudence' (von Stein) – a mixture of both formal budget and tax 'principles' – and a highly pedantic and descriptive systematization of facts and definitions. Analytical economics, insights into the laws of the market and the study of the interaction between market and state (or even of the bureaucratic and political mechanism) are relatively unknown in the simple textbooks of the cameralists, which show otherwise sound common sense. Statistics, important for census and grasping foreign trade, became a new discipline of the cameral curriculum.

The *practical* policy of cameralism concentrated on the development of a country which had been devastated and depopulated in the 30 Years' War and impoverished by the discovery of the sea route to India and the fall of Constantinople. Under these abnormal circumstances a political and bureaucratic monopoly attempted to reconstruct the economic foundations of the country by an active population policy, the establishment of state manufactures and banks, the extension of infrastructure (canals, bridges, harbours and roads) and the promotion of modernization. It strictly regulated the still important agricultural sector, as well as trade and commerce.

The state protected the trades (*Gewerbe*) by means of high tariffs to restrict imports of unnecessary raw materials and it facilitated exports of manufactures and import substitution. On the other hand, the government removed internal trade barriers by abolishing the medieval guild organization and by unifying the law for municipalities. Mercantilist efforts to augment the state treasure via trade surplus and money policy were, of course, another main cameralistic aim. Finally, it is notable that its monetary policy was inconsistent, in so far as the hoarding of precious metals as opposed to their circulating function was not clearly distinguished.

To set cameralism in secular perspective, the famous arguments of Smith and the Physiocrats against the 'mercantile system' seem to be *mutatis mutandis* valid for neo-mercantilism, which also justifies both state intervention in the market and a greater GNP government share and often reverts to the regulatory rules and the principles of planning in this former epoch. However, neo-mercantilism fails to prove seriously both the state's competence to ensure efficiency and equity in the public sector and its ability to regulate the market reasonably. Some writers tend to overlook that in our times the basic conditions in the state and the economy are radically different from those of three centuries ago. For example, economic, political and administrative conditions in the German principalities differed strikingly from Ludwig Erhard's situation after the Second World War. And the wide gap between the Great Depression of the 1930s and the technologically influenced stagflation of the 1980s was obviously so fundamental that the regulatory Keynesian budget and employment theory, with its then unrealistic assumptions, became rather obsolete. Thus any attempt to revive the strict regulating prescriptions of all-embracing cameralism, which lacks sufficient analysis and empirical testing, would apparently be a violation of both reason and experience. In this case we would use analytically poor (and old) tools to repair the wrong (and modern) machine.

Bibliography

Becher, J.J. 1668. *Politischer Diskurs*. Frankfurt/Main: Bielcke.

Humpert, M. 1935–7. *Bibliographie der Kameralwissenschaften.* Cologne: Schroeder. (Includes nearly 14,000 items.)

Mohl, R. 1855–8. *Geschichte und Literatur der Staatswissenschaften.* vols. 1–3. Erlangen: Enke.

Recktenwald, H.C., ed. 1973. *Political economy a historical perspective*. London: Collier–Macmillan.

Recktenwald, H.C. 1986. *Das Selbsinteresse – Zentrales Axiom der ökonomischen Wissenschaft*. Wiesbaden: Leibniz-Akademie der Wissenschaften.

Small, A.W. 1909. *The cameralists*. New York: Franklin.

Sommer, L. 1920–5. *Die österreichischen Kameralisten in dogmengeschichtlicher Darstellung*, 2 vols. Vienna: Konegen.
von Justi, J.H.G. 1755. *Staatswirtschaft*. Leipzig: Breitkopf.
von Sonnenfels, J. 1765–76. *Grundsätze der Polizei, Handlungs- und Finanzwissenschaft*. Vienna: Camesina.

Campaign Finance, Economics of

Scott Ashworth

Abstract

This article surveys recent work aimed at evaluating the welfare effects of campaign finance reform. The theoretical literature distinguishes two types of contributor: those who desire ideological policies and those who want personal favours. A series of models shows that these different types of contributor have different implications for campaign finance regulation. The models also give some suggestions about the sort of empirical evidence that would argue for or against certain campaign finance regulations. These suggestions have been followed up by recent empirical work.

Keywords

Advertising; Campaign finance reform; Campaign finance, economics of; Election; Incumbency; Matching funds; Multiple equilibria; Probabilistic voting; Voting

JEL Classifications

D71

Campaign finance is a contentious issue in American politics. Reformers charge that a system in which interest groups provide the funds for campaigns creates opportunities for corruption, while others argue that restrictions on donations would limit the provision of information to voters. For an economist, the natural way to evaluate such arguments is to construct a model that explicitly treats the preferences and beliefs of the voters, to deduce the conditions under which the model predicts welfare improvements from regulation, and to check empirically if these conditions hold in actual elections. This article surveys a recent body of literature that does just that.

First-Generation Models

Early work on campaign finance took a reduced-form approach to the link between campaign activity and votes (Austen-Smith 1987; Baron 1989, 1994; Grossman and Helpman 1996; Snyder 1990). This literature identified two ideal types of contributor: position-induced contributors, who help ideologically compatible candidates win office, and service-induced contributors, whose contributions are analogous to purchasing contingent claims on favours provided to the buyer at the expense of citizens in general.

This literature yielded several important insights. For example, Baron (1989) finds that trades of contributions for promises of favours have interesting implications for the incumbency advantage (see, for example, Gelman and King 1990, and Ansolabehere and Snyder 2002, for empirical work on the incumbency advantage in US elections). A candidate with an exogenous advantage is more likely to be able to deliver the promised favours, making the promise more valuable. Thus an advantaged candidate can raise funds on more favourable terms, reinforcing the advantage. Morton and Myerson (1992) show that this mechanism can even lead to multiple equilibria, where predictions that one candidate will win become self-fulfilling because contributions flow to the presumptive winner.

As the comprehensive survey of this literature by Morton and Cameron (1992) emphasizes, this approach cannot address the welfare qsts raised by proposals for campaign finance reform. We now turn to more recent research that ‘opens up the black box’ and provides some welfare analysis.

Microfounded Models

A bare-bones model illustrates the main points of the literature. The game has four players: two candidates, a voter, and an interest group.

Each candidate has some level of 'quality', which could be either ability or ideological similarity to the voter. The key is that quality is valued by the voter. Candidate i's ability is θ_i. It is common knowledge that $\theta_1 = 1$, and that θ_2 is equally likely to be 0 or 2. Each candidate maximizes his probability of winning.

At the start of the game, the candidates learn θ_2, but the voter does not. At cost $c \in (0, 1)$, candidate 2 can truthfully reveal θ_2. Candidates have no funds of their own. The interest group has sufficient funds to pay for the information transmission, if it wants to.

Even without specifying the group's payoffs, we can derive two benchmarks. *The no-campaign solution*. First, assume the interest group is prohibited from funding candidate 2's campaign. Then the voter goes to the polls not knowing θ_2. Thus she is indifferent between the two candidates, and gets expected payoff 1 no matter how she votes. The natural voting rule is to have her toss a fair coin. (This would be the outcome if there were a mean-zero popularity shock prior to the election.) In this case, each candidate gets payoff 1/2.

The voter's optimum. Second, assume there is a planner who can observe the true θ_2 and communicate it to the voter, paying for the communication with a lumpsum tax on the voter.

Announcing the true θ in only one of the states suffices for complete communication, and allows for a cost savings compared with always announcing the state. So the planner announces θ_2 if and only if $\theta_2 = 2$, and the voter votes for 2 if there is an announcement and for 1 if not. Her payoff is

$$\frac{1}{2} + \frac{1}{2}(2 - c) = \frac{3}{2} - \frac{1}{2}c > 1.$$

Thus the voter is better off than in the no-campaign solution. Furthermore, each candidate still wins with *ex ante* probability 1/2, so the policy represents an *ex ante* Pareto improvement over the no campaign solution.

This scheme would be hard to implement, because it is vulnerable to collusion between the regulator and candidate 1. Thus we are interested in whether or not interest-group finance can improve on the no-campaign benchmark.

Position-Induced Contributors

Now assume the interest group wants candidate 2 to win independent of θ, perhaps because it shares the candidate's ideology. Formally, the group's payoff is

$$bw - k,$$

where $b > 0$ is the payoff to the group from having 2 win, w is an indicator variable equalling 1 if and only if candidate 2 wins, and k is the contribution to candidate 2. The timing is:

1. The candidates and the group learn θ_2.
2. The group chooses a contribution $k > 0$.
3. If $k \geq c$, the candidate decides whether or not to advertise θ.
4. The voter sees any ads purchased, and then selects the winner.

Proposition 1 *If $b > c$, then there is a perfect Bayesian equilibrium (PBE) in which*

- *the group contributes c if and only if $\theta_2 = 2$* and
- *the voter chooses candidate 2 if and only if she sees an ad certifying that $\theta_2 = 2$.*

The idea is simple. The group is better off if 2 wins. If $\theta_2 = 2$, the group can ensure that 2 wins by funding a campaign informing the voter of her true preference for 2. And if the benefit from having 2 win (b) exceeds the cost (c), the group wants to do this. Finally, the group does not contribute to a low type of candidate 2 – this cannot help the group because the candidate cannot lie.

If there are contributions in equilibrium, then the voter gains over the no-campaign solution, having a payoff of $3/2 > 1$. Thus banning contributions reduces the voter's welfare. Furthermore, the equilibrium without contributions is Pareto dominated by the following matching fund policy. Fix γ strictly between 0 and b. If the group donates γ to candidate 2, then the regulator kicks in $c - \gamma$, paid for by a lump-sum tax on the voter. The group's *ex ante* payoff increases from 0 to $(b - \gamma)/2$ and the voter's payoff increases from 1 to $3/2 - (c - \gamma)/2 > 1$. The candidates are indifferent at the *ex ante* stage.

Coate (2003) elaborates on this story in two ways. First, the voter is uncertain about both ideologies, and both candidates can receive contributions. Second, and more importantly, candidates are selected by the party's median member, who has different preferences from the median in the electorate. (Here quality is the inverse of distance from the median.) The interest group prefers less moderate candidates. However, the groups prefer to fund more moderate candidates – campaign ads are effective only when the ad reveals that the candidate is more moderate than a non-advertising candidate. This gives the party an additional incentive to choose moderate candidates, because moderate candidates can raise funds and thus do well in the election. In equilibrium, the party mixes between moderate and extremist candidates.

In this environment, simply banning contributions creates both winners and losers. Moderate voters lose. First, they must make their choices with worse information, as in the bare-bones model above. Second, candidates are less likely to be moderate. Members of the interest groups, on the other hand, are better off. They save the cost of the contributions, and policy is no worse in expectation – the extra probability that policy is extreme in the wrong direction is exactly offset by the increased probability that policy is close to the group.

Service-Induced Contributors

Now assume the group does not care directly who wins the election. Instead, the group values transfers from the winner. The group and candidate 2 can sign a contract specifying that candidate 2 receives c from the group, and, if he wins, he transfers the amount t to the group. This transfer if financed by a tax on the voter of $(1 + \lambda)t$, where λ represents the deadweight loss of the transfer.

The timing is:

1. The candidates and the group learn θ_2.
2. Candidate 2 makes a take it or leave it offer of a contract t to the group.
3. The group accepts or not.
4. If the contract is accepted, the candidate decides whether or not to advertise θ.
5. The voter sees any ads purchased, and then selects the winner.

Proposition 2 *If* $(1 + \lambda)c \leq 1$*, then there is a PBE in which the group funds the campaign if and only if* $\theta_2 = 2$ *and the voter selects candidate 2 if and only if she sees an ad certifying* $\theta_2 = 2$.

Again, the basic idea is simple. If the voter sees an ad, she learns two things. First, she learns that $\theta_2 = 2$, which improves her evaluation of candidate 2. Second, she learns that the group and the candidate have made a deal, so electing candidate 2 costs her $(1 + \lambda)c$. This tradeoff is acceptable if $(1 + \lambda)c \leq 1$.

In such an equilibrium, the voter's payoff is

$$\frac{1}{2} + \frac{1}{2}(2 - (1 + \lambda)c) = \frac{3}{2} - \frac{(1 + \lambda)c}{2}.$$

This payoff is lower than the voter-optimal benchmark payoff by $\lambda c/2$.

Again, matching funds can help. Assume again that the regulator pays $c - \gamma$ of the cost. This policy reduces the welfare loss compared with the benchmark to $\lambda\gamma/2 < \lambda c/2$.

Most papers in the literature introduce some uncertainty in the voting stage. With this addition, Prat (2002), Coate (2004) and Ashworth (2006) show that the candidate might promise so much that the voter actually *loses* from the campaign. To see the intuition, consider the candidate's incentive to advertise. Without probabilistic voting, the incentive to expand transfers is limited – once the voter's cost of transfers passes 1, the probability of election changes discontinuously from 1 to

0. With probabilistic voting, by contrast, small changes in transfers have similarly small effects on the re-election probability. In this case, candidates have an incentive to expand transfers all the way to the point where the voter is indifferent between a high-quality candidate with transfers and a low-quality candidate with no transfers. In such a case, the voter actually loses from the possibility of a campaign, and would be better off if contributions were banned outright – the likelihood of getting a high-quality winner is no lower, and the voter escapes the cost of favours.

The key to the inefficiency here is that the voter's knowledge that ads imply favours to interest groups makes the ads less effective at ensuring a high-quality candidate is elected.

Again, matching funds might be a better solution. In Coate (2004), the scale of the campaign can vary continuously. Greater spending increases the fraction of the (large) electorate that is informed. Matching funds come into play if the benefit from winning is low enough that ads are not rendered totally ineffective. In that case, a limit on contributions reduces the amount of favours, preserving the effectiveness of the ads. And the matching funds allow the scale of the campaign to be unchanged from the unregulated case.

So far, matching funds have seemed like a great policy. But they have a cost in asymmetric contests. In Ashworth (2006), the scale of campaigns is fixed (as in the bare-bones model above), but candidate 2 has an advantage independent of advertising. For moderate levels of the advantage, the advantaged candidate mounts a costly campaign even though the value of the information to the voter is less than the cost the voter pays *ex post*. For greater values of the advantage, no campaign takes place in equilibrium – the possible increase in the voter's evaluation is too small to outweigh the promised favours. Matching funds can increase the likelihood of an active campaign in such cases, even though reducing their likelihood would be efficient.

Hard vs. Soft Information

The literature focuses on two mechanisms that make advertisements informative. The first is the one we have relied on above, namely, the candidate may have verifiable information, information that cannot be falsified. The second, studied by Gerber (1996), Prat (2002), and Potters et al. (1997), is indirectly informative campaigns. Interest groups observe the quality of the candidates, but voters do not. If groups condition their contributions on quality, then voters can learn about quality by inverting the contribution schedule. Gerber and Prat show that equilibria with informative advertising exist, even thought the ads have no direct informational content. As in the case with hard information, service-induced contributions imply that a ban on contributions can benefit the median voter. On the other hand, public financing would have no value with indirectly informative advertising – there's no signal if the election regulator hands out funds to everyone. Thus a non-trivial policy problem of public financing arises only with directly informative advertising.

Empirics

Do Contributions Buy Favours?

Contributors' motivations played a key role in the welfare conclusions above. What do the data say about these motivations? The most direct approach to this question looks at correlations between donations from interest groups and votes that those groups care about. For example, we could regress votes in favour of increasing the minimum wage on contributions from unions. Of course, a positive correlation on its own does not discriminate between the theories – are the union contributions changing votes or do unions just contribute to exogenously union-friendly candidates? The many studies that try to disentangle these forces affecting roll-call votes find only weak evidence that contributions buy votes (Ansolabehere et al. 2003). One interpretation is that contributions are positioninduced rather than service-induced.

However, focusing on roll calls misses much Congressional activity (Hall 1996). Thus researchers have also looked to more indirect evidence. For example, Gordon and Hafer (2005) find that firms making large donations are less

monitored by agencies, suggesting that donations induce members of Congress to interfere in regulatory oversight. Many papers have shown that political action committees (PACs) direct their contributions in ways more consistent with service-induced motivations than with position-induced motivations (Kroszner and Stratmann 1998; Romer and Snyder 1994; Snyder 1990). Perhaps the most convincing is McCarty and Rothenberg (1996), who document that individual PACs made significant shifts in donations from Democrats to Republicans after the Republicans took control of Congress in 1994, suggesting that the contributions were not ideological.

Attempts to directly estimate the impact of contributions on policy have not reached a consensus, except that the effects are smaller than public outcry might suggest (Ansolabehere et al. 2003). The next subsection turns to a more theory-driven approach to evaluating the potential for welfare gains from regulation.

Spending and Election Outcomes

A substantial empirical literature has tried to estimate the effect of campaign spending on electoral outcomes. Cross-sectional analyses that do not condition on incumbent quality show that challenger spending is associated with better electoral performance, but incumbent spending is unrelated to success. (See the discussion in Jacobson 2001, ch. 3, which summarizes the extensive empirical work initiated by Jacobson 1978.) Of course, interpreting these correlations is difficult because of an endogeneity problem – candidates spend more when they expect the race to be competitive. Several researchers have tried to deal with this endogeneity issue (Green and Kranso 1988; Levitt 1994; Gerber 1998; Erikson and Palfrey 1998; 2000). These papers all find that spending is roughly equally effective for both incumbents and challengers, but there is no consensus about the size of the effects. (Looking across several of the most prominent estimates, Gerber 2004, calculates an implied cost for a House incumbent to get one additional vote ranging from $15 to $367.)

Prat (2000) points out that, even when one controls for candidate quality, there is an identification problem in these regressions. Simply put, the functional relationship between spending and election outcomes (with quality held fixed) depends on the way funds are raised. To see this, consider the models of service-induced contributions discussed previously. In all of the models, an exogenous increase in quality has two effects. First, the candidate raises more funds and informs the voters of his high quality, which helps his electoral chances. Second, the voter infers that the funds were given in exchange for promises of favours, which hurts his electoral chances. Thus the regressions estimate 'the effect on electoral outcomes of an extra dollar of campaign spending net of the political cost of persuading lobbies to donate the extra dollar' (Prat 2006, p. 60).

In addition to providing an important critique of the standard inpts of the empirical evidence, the prediction that the effectiveness of advertising is decreasing in the degree of service-induced contributing provides a way to test empirically for the possibility of welfare-improving policy. In particular, the theoretical models suggest that limits on contributions and (perhaps) matching funds can improve welfare precisely when campaign spending is ineffective. Thus the prediction of reduced effectiveness speaks directly to the welfare implications of the models.

Stratmann and colleagues have been leaders in testing these implications. Houser and Stratmann (2006) carry out laboratory experiments modelled after the theoretical set-up of Coate (2004) and Ashworth (2006). High-quality candidates are more likely to win in a public financing treatment than in a privately financed treatment. They also find that margins of victory are greater in the public financing treatment. In a treatment with caps on contributions, they find that voter welfare goes up, but the probability of electing a high-quality incumbent does not. These experiments support the theoretical predictions, suggesting that voters are capable of inferring that interest-group financed ads imply that the candidate has promised favours.

Stratmann (2006) exploits state-level variation in campaign finance laws to see whether the theoretical predictions hold up in field data. He first estimates standard vote-share/spending

regressions for each state's House elections. He then examines the relationship between the effectiveness of spending and the existence of limits on contributions. As predicted by the theory, he finds that effectiveness is lower when campaign finance regulations are more liberal. These results hold for all of incumbents, challengers, and open-seat candidates. Stratmann and Aparicio-Castillo (2006) show that states that limit giving subsequently have lower incumbent vote shares. This finding is consistent with Baron's (1989) and Ashworth's (2006) theoretical finding that the financing process can exaggerate incumbency advantages.

See Also

- ▶ Political Competition
- ▶ Political Institutions, Economic Approaches to
- ▶ Rent Seeking

Bibliography

Ansolabehere, S., and J.M. Snyder Jr. 2002. The incumbency advantage in U.S. elections: An analysis of state and federal offices, 1942–2000. *Election Law Journal* 1: 315–338.

Ansolabehere, S., J.M. de Figueiedo, and J.M. Snyder Jr. 2003. Why is there so little money in U.S. politics? *Journal of Economic Perspectives* 17(1): 105–130.

Ashworth, S. 2006. Campaign finance and voter welfare with entrenched incumbents. *American Political Science Review* 100: 55–68.

Austen-Smith, D. 1987. Interest groups, campaign contributions, and probabilistic voting. *Public Choice* 54: 123–139.

Baron, D.P. 1989. Service-induced campaign contributions and the electoral equilibrium. *Quarterly Journal of Economics* 104: 45–72.

Baron, D.P. 1994. Electoral competition with informed and uninformed voters. *American Political Science Review* 88: 33–47.

Coate, S. 2003. Political competition with campaign contributions and informative advertising. *Journal of the European Economic Association* 2: 772–804.

Coate, S. 2004. Pareto-improving campaign finance policy. *American Economic Review* 94: 628–655.

Erikson, R.S., and T.R. Palfrey. 1998. Campaign spending and incumbency: An alternative simultaneous equations approach. *Journal of Politics* 60: 355–373.

Erikson, R.S., and T.R. Palfrey. 2000. Equilibria in campaign spending games: Theory and data. *American Political Science Review* 94: 595–609.

Gelman, A., and G. King. 1990. Estimating incumbency advantage without bias. *American Journal of Political Science* 34: 1142–1164.

Gerber, A. 1996. *Rational voters, candidate spending, and incomplete information: A theoretical analysis with implications for campaign finance reform*, Working Paper No. 96–01.1, Institution for Social and Policy Studies, Yale University.

Gerber, A. 1998. Estimating the effect of campaign spending on senate election outcomes using instrumental variables. *American Political Science Review* 92: 401–411.

Gerber, A.S. 2004. Does campaign spending work?: Field experiments provide evidence and suggest new theory. *American Behavioral Scientist* 47: 541–574.

Gordon, S.C., and C. Hafer. 2005. Flexing muscle: Corporate political expenditures as signals to the bureaucracy. *American Political Science Review* 99: 245–261.

Green, D.P., and J.S. Kranso. 1988. Salvation for the spendthrift incumbent: Reestimating the effects of campaign spending in House elections. *American Journal of Political Science* 32: 884–907.

Grossman, G.M., and E. Helpman. 1996. Electoral competition and special interest politics. *Review of Economic Studies* 63: 265–286.

Hall, R.L. 1996. *Participation in Congress*. New Haven: Yale University Press.

Houser, D., and T. Stratmann. 2006. *Selling favors in the lab: Experiments on campaign finance reform*, Working Paper No. 1727, CESifo.

Jacobson, G.C. 1978. The effects of campaign spending in Congressional elections. *American Political Science Review* 72: 469–491.

Jacobson, G.C. 2001. *The politics of congressional elections*, 5th ed. New York: Addison-Wesley Educational Publishers, Inc.

Kroszner, R.S., and T. Stratmann. 1998. Interest Group competition and the organization of congress: Theory and evidence from financial services political action committees. *American Economic Review* 88: 1163–1187.

Levitt, S.D. 1994. Using repeat challengers to estimate the effect of campaign spending on election outcomes in the U.S. House. *Journal of Political Economy* 102: 777–798.

McCarty, N., and L.S. Rothenberg. 1996. *Investment in politicians? Evidence from the 1994 elections*. Paper presented at the 1996 annual meeting of the American Political Science Association, 29 August–1 September.

Morton, R., and C. Cameron. 1992. Elections and the theory of campaign contributions: A survey and critical analysis. *Economics and Politics* 4: 79–108.

Morton, R.B., and R.B. Myerson. 1992. *Decisiveness of contributors' perceptions in elections*. Working paper, New York University.

Potters, J., R. Sloof, and F. van Winden. 1997. Campaign expenditures, contributions, and direct endorsements: The strategic use of information to influence voter

behavior. *European Journal of Political Economy* 13: 1–31.
Prat, A. 2000. Campaign spending with office-seeking politicians, rational voters, and multiple lobbies. *Journal of Economic Theory* 103: 162–189.
Prat, A. 2002. Campaign advertising and voter welfare. *Review of Economic Studies* 69: 997–1017.
Prat, A. 2006. Rational voters and political advertising. In *The Oxford handbook of political economy*, ed. B.R. Weingast and D. Wittman. Oxford: Oxford University Press.
Romer, T., and J.M. Snyder Jr. 1994. An empirical investigation of the dynamics of PAC contributions. *American Journal of Political Science* 38: 745–769.
Snyder, J.M. 1990. Campaign contributions as investments: The U.S. House of Representatives, 1980–1986. *Journal of Political Economy* 98: 1195–1227.
Stratmann, T. 2006. Contribution limits and the effectiveness of campaign spending. *Public Choice* 129: 461–474.
Stratmann, T., and F.J. Aparicio-Castillo. 2006. Competition policy for elections: Do campaign contribution limits matter? *Public Choice* 127: 177–206.

Canada, Economics in

Robert W. Dimand and Robin F. Neill

Abstract

In the first half of the 20th century, economics in Canada was primarily economic history, and its contribution was the staple theory of Canadian economic development. After the Second World War Keynesian macroeconomics swept the nation and, despite its British origin, it indigenized into a theory of primary product export-based growth, and a Western Marxist theory of the staple trap. In the last quarter of the century, positivism, monetarism, and neoconservative new classical economics swept north from the United States, leaving only the specific domestic circumstances to which it was applied as the distinctive thing about economics in Canada.

Keywords

Ashley, W. J.; Bladen, V.; Brittnell, G.; Bryce, R.; Canada, economics in; Conspicuous consumption; Cournot, A. A.; Dales, J.; Dependency; Easterbrook, W.; Ely, R. T.; English Historical School; Fowke, V.; Galbraith, J. K.; German Historical School; Gourlay, R.; Hume, D.; Infant-industry protection; Ingram, J.K.; Innis, H. A.; Institutional economics; Internal rate of return; Johnson, H. G.; Mackintosh, W. A.; Mathematical methods in political economy; Mavor, J.; Monetary policy rules; Mundell, R.; Periphery; Plumptre, A. F. W.; Political arithmetic; Rae, J. (1796–1872); Saunders, S.; Social Credit; Staples thesis; Technical change; Time preference; Timlin, M. F.; Viner, J.; Wakefield, E. G

JEL Classifications

B1

The pre-history of economics in Canada begins with the description of the society and products of New France by Pierre Boucher (1664), a former governor at Trois Rivieres and the founding seigneur of Boucherville, writing in the political arithmetic tradition of Boisguilbert and Vauban. The most notable of such descriptive works was, after the British conquest, the vast, disorganized, but often incisive *Statistical Account of Upper Canada* by the political dissident Robert Gourlay (1822), whose criticism of unrepresentative and corrupt government led to his exile as an undesirable alien - on the grounds of his birth in Scotland rather than England (Dimand 1992). Although, apart from Boucher and Gourlay, early descriptive writings about settlement and economic conditions in Canada tended to have little economic analysis, Boucher displayed an intuitive sense of economies of scale, urging that policy should encourage concentration of settlement in small areas, where mutually beneficial exchange would lead to a surplus product. Independently, Gourlay later formulated a linear relationship between land values and the number of inhabitants per acre. He urged the government to borrow to fund increased immigration and settlement, paying off the loan by taxing the resulting increase in land value. The influence of Gourlay's theorizing about the appropriate structure of property rights to promote population density in a newly settled colony (such as

limiting the size of land grants to avoid dispersion of settlers) was acknowledged by Edward Gibbon Wakefield, the English theorist of colonization who wrote Appendix B on land policy for Lord Durham's report on Canada after the 1837 rebellion and then served in the Canadian legislature before taking a leading role in the settlement of New Zealand (Wakefield 1968; Goodwin 1961, ch. 1; Neil 1991, ch. 1). One Canadian topic, the playing card currency of New France, so often cited by later economic historians (for example, Shortt 1987), attracted the attention of one of the great early economists, the philosopher David Hume, as British charge d'affaires in Paris after the Seven Years War and then as Under-Secretary of State; Hume negotiated the settlement of the outstanding paper money of New France after the British Conquest (Dimand 2005).

John Rae was an outstanding 19th-century economic theorist who wrote his *New Principles of Political Economy* (1834) while headmaster of the Gore District Grammar School in Upper Canada (now Ontario). Rae, although born and educated in medicine in Scotland, eventually became a district judge in the Kingdom of Hawaii before dying in Staten Island. For decades, he was known primarily through John Stuart Mill's citation of his statement of the infant industry argument for protection: although Sir John A. MacDonald, Canada's first prime minister, cited Rae in support of his national policy of tariff protection for manufacturing, he seems to have known of Rae only through Mill (MacDonald, quoted in Neill 1991, pp. 85–91). C.W. Mixter's new, rearranged edition of Rae's book in 1905 revealed Rae's analysis of 'effective desire of accumulation' as a pioneering capital theory, and two years later Irving Fisher dedicated The Rate of Interest 'to the memory of John Rae who laid the foundations upon which I have endeavored to build', acknowledging Rae for foreshadowing both time preference and internal rate of return over costs. Rae has since been celebrated for his discussions of conspicuous consumption, more than six decades before Thorstein Veblen, and of endogenous technical change (James 1965; Hamouda et al. 1998). University of Toronto mathematics professor John Bradford Cherriman (educated at St John's College, Cambridge, a few years before Alfred Marshall) made another striking, but isolated, contribution to economic theory: a ten-page review article and exposition of Cournot's essay in mathematical economics of 19 years before, endorsing the mathematical approach to political economy, hailing Cournot's work as more important than Ricardo, and long antedating Joseph Bertrand's 1883 article that was long thought to be the first review of Cournot's 1838 volume (Cherriman 1857; Dimand 1988, 1995). More characteristic of this period than the theorizing of Rae and Cherriman were the numerous practical and descriptive discussions of economic affairs, economics in the context of action (see Goodwin 1961; Neil 1991; Neill and Paquet 1993).

The Rise of Academic Economics in Canada

Although a few courses had been offered previously, economics in Canadian universities began in 1888 with the appointment of the English historical economist W.J. (later Sir William) Ashley as professor of political economy and constitutional history at the University of Toronto and of Adam Shortt (previously tutor in philosophy, instructor in botany, and demonstrator in chemistry) as lecturer in political economy at Queen's University, Kingston (promoted to Sir John A. Macdonald Professor of Political Science in 1891). Professorial appointments at the university were then made by Order in Council by the provincial government, and candidates were interviewed by the Premier of Ontario and by the chancellor of the University. No classical or neoclassical theorist would have been appointed, lest they promote free trade in their lectures, but the English Historical School was acceptable (Drummond 1983). When Ashley departed in 1892 to become professor of economic history at Harvard (and later dean of commerce in Birmingham), he was succeeded by James Mavor, Scottish economic historian of Russia and friend to Tolstoy, Kropotkin, and the Doukhobors (see Mavor 1923), and until 1970 the Department of Political Economy was led by a succession of distinguished

economic historians (apart from one sociologist), notably Harold Innis and William Easterbrook, and the historian of economic thought Vincent Bladen (see Drummond 1983; Bladen 1978). Under Ashley's sponsorship, the University of Toronto published the first academic economic writing by a Canadian woman, Jean Scott Thomas (1889), 'The conditions of female labour in Ontario'. As in other disciplines and elsewhere in the Dominions and the British Empire, several early professors of economics in English-speaking Canadian universities, notably Ashley and C.R. Fay in Toronto and A.W. Flux at McGill, were British scholars who had finished their careers in Britain, as was James Bonar, Deputy Master of the Mint in Ottawa and authority on Malthus. The British Association for the Advancement of Science met in Montreal in 1884; in other years it met in Dublin, Cape Town, or Sydney. The following year, the association commemorated its meeting with *Canadian Economics,* a volume of 27 papers by Canadian and American authors that, according to Goodwin (1961, p. 116), 'marked the end of an era when description and analysis were carried out by interested persons in all walks of life and before there were any professional economists in government and the universities' . The Canadian Political Science Association met in September 1913, with Adam Shortt as president, and published a volume of proceedings, but the September 1914 meeting was cancelled when the First World War broke out, and the association lapsed until 1929.

Long after the social sciences separated in Britain and the United States, they remained institutionally linked in Canada, sharing a single Department of Political Economy at the University of Toronto until 1982 (the equivalent term at McGill and the University of Saskatchewan was Department of Economics and Political Science), a single Canadian Political Science Association and the *Canadian Journal of Economics and Political Science* (first published in 1935) until 1966 (the sociologists and anthropologists seceded in 1963), with the economists departing only much later from the joint annual conferences of the Learned Societies (now the Humanities and Social Science Congress). As Taylor (1960, p. 8) remarks, 'Shortt, Skelton, Mavor, and Leacock throughout their careers could almost equally well be described as historians or political scientists.' While the economic historian Harold Innis headed Toronto's Department of Political Economy during the 1930s and 1940s, scholars in the various disciplines there, not all of them within the department, were linked by their historical approach and by Innis's influence, in historical sociology (S.D. Clark), history of political thought (C.B. Macpherson), history of economic thought (Vincent Bladen), economic history (John Dales, William Easterbrook), historical geography (Andrew Hill Clark), history of communications (Marshall McLuhan), Canadian history (Donald Creighton, Innis's biographer). Formal economic theory, in contrast, was conspicuously absent, except that A.F.W. Plumptre, before joining the public service, taught Keynes's *Treatise on Money,* having studied in Cambridge while that book was being written. When the University of Saskatchewan opened in 1910, economics was taught by the professor of history, using texts by Richard T. Ely, an American economist influenced by the German Historical School, and by Ashley, Archdeacon William Cunningham, and J. Kell Ingram of the English Historical School, but not Marshall or Jevons (Spafford 2000). One consequence of multidisciplinary sharing of departments, association, and journal was that after the humorist Stephen Leacock, trained in political science and author of a successful textbook in that field, succeeded Flux as Dow Professor of Economics and Political Science at McGill in 1908, he acquired public credibility for his economic pronouncements, such as advocating a tariff union for the British Empire to end the Great Depression.

Growing numbers of academics, and the gains from division of labour in scholarly research and publication as in other activities, led the social sciences in Canada to become increasingly separate after the Second World War, well in advance of formal institutional separation. The British connection and the emphasis on a historical approach also faded in the same decades, as Canadian economics became more grounded in formal theory

and quantitative methods and more attuned to intellectual developments in the United States.

The teaching of economics emerged later in French Canada. The journalist Etienne Parent (1846), an admirer of Adam Smith and Jean-Baptiste Say, was unusual in declaring political economy a science and urging the enlightened publication of the principles it taught, notably free trade and the respectability of commerce and industry as occupations. Although Parent became Under-Secretary of State when the Dominion of Canada was created in 1867, his views on the study of economics had little influence. Political economy was widely identified with doctrinaire free traders (such as Parent) and with the secular pursuit of material gain, and did not often find a place in the curriculum of the Jesuit classical colleges in Quebec, which steered promising students towards law, medicine and the Church. Attitudes toward social and economic research in Quebec changed following papal social encyclicals such as *Rerum Novarum* in 1891 (an influence that ceased to dominate intellectual life in Quebec after the 1960s). The École des Hautes Études Commerciales (HEC) was established in Montreal in 1911, and its journal *Actualité Ėconomique* began publication in 1925. Such HEC professors as Esdras Minville (1979), Edouard Montpetit (1939–42), and François-Albert Angers were concerned with the economic independence and distinctive cultural values of French Canadian society, beyond the technical aspects of the economics that Montpetit had studied under Charles Gide at the Sciences-Po in Paris, and the concerns of French Canadian economists were shaped by the uneasy relationship of their intellectual milieu and society with the rest of Canada and North America (see Falardeau 1944; Angers 1961; Parizeau 1968; and the extensive oral history in Paquet 1989 on the emergence and evolution of francophone economics in Canada).

The Staples Thesis

The two outstanding figures of inter-war Canadian economics, William A. Mackintosh (1923, 1939), of Queen's University, and Harold A. Innis (1930, 1940, 1956), of the University of Toronto, developed a distinctive approach to understanding Canada's economic development, the staples thesis (see also Mary Quayle Innis 1935; Creighton 1937; Neill 1972). Rejecting the universal applicability of neoclassical analysis of the market determination of relative prices, the staples thesis drew on a wide range of influences (including American institutionalists, notably Veblen) to argue that a newly settled, peripheral economy could not be studied in the same way as the core economies of the world economy. The keys to analysing Canadian economic development were the geographical setting (especially regional differences and the transport routes such as the St Lawrence Valley/Great Lakes) and the characteristics of the staple commodities such as cod, fur and wheat that successively dominated an export-oriented peripheral economy. The core-periphery distinction in the staples thesis was mirrored in the structure of interwar Canadian economics discipline: Mackintosh and Innis at the leading universities in the industrial and commercial heartland of Ontario developed the dominant interpretation of Canadian development as whole, while George Brittnell (1939) and Vernon Fowke (1946) at the University of Saskatchewan focused on the locally dominant staple, wheat, and maritime economists such as Stanley Saunders (1939) were concerned with the maritime provinces as an economically backward region within Confederation. This historical and institutional approach, which had parallels in later Latin American dependency theory, received considerable attention beyond Canada: at the time of his death in 1952, Innis had been elected president of the American Economic Association, the only foreigner or non-resident ever so honoured. Except for Creighton on the merchant class, the staple literature paid little attention to class until H. Clare Pentland's Toronto dissertation on the emergence of Canada's industrial working class, finished in 1961 and published posthumously 20 years later, but largely written at the University of Toronto before Innis's death (Pentland 1950, 1981). Canadian political economy influenced by Innis and Pentland continues to flourish in the disciplines of political science and sociology (and Innis 1951, is influential in

communications studies in Canada), but has largely disappeared from economics departments, as Canadian economics has become part of an international mainstream in which the old (or original) institutional economics, widespread in the interwar United States, has been marginalized.

Economists in and on Government in Canada

The Dominion Bureau of Statistics (now Statistics Canada) became a leading centre of quantitative research under Robert Coats, for 25 years the first Dominion Statistician, an achievement recognized internationally by the election of Coats as president of the American Statistical Association in 1938 (see Coats 1932; Keyfritz and Greenway 1961). Economists at Queen's and McMaster Universities produced two volumes of *Statistical Contributions to Canadian Economic History* in 1931. Economists became deeply involved in other areas of government, more so than in many other countries. After exploring Canada's monetary and banking history in a long series of articles in the *Journal of the Canadian Bankers Association* (reprinted as Shortt 1987), Adam Shortt, the first economics professor at Queen's University, came to Ottawa to head the Civil Service Commission and then to superintend the publication of numerous documents on monetary history (see Shortt 1976). His student and successor at Queen's, Oscar D. Skelton, winner of the Hart Shaffner & Marx Prize for a study of socialism (Skelton 1911), was Under-Secretary of State for External Affairs from 1925 until his death in 1941, an especially important position because the External Affairs portfolio was held by the prime minister, so that Skelton was the prime minister's deputy minister. Skelton in turn recruited another Queen's economics professor, W. Clifford Clark, as Deputy Minister of Finance from 1932 until Clark's death in 1952. Noteworthy anniversary surveys of the progress of economic scholarship in Canada were written by the Under-Secretary of State for External Affairs (Skelton 1932) and the Deputy Minister of Finance (Taylor 1960), rather than by academics, and economic research in Quebec was surveyed by a future separatist Finance Minister and Premier of Quebec (Parizeau 1968).

The Great Depression of the 1930s, which was especially severe in the Prairie provinces, and the Second World War expanded the role of the government in the economy, and of economists in government, notably with the creation of the Bank of Canada in 1934 and of a system of national accounts during the war. The extent of popular dissatisfaction with existing economic arrangements was shown in 1935 when Alberta gave 56 of the 63 seats in its provincial legislature (and, later that year, all 15 of its seats in the federal House of Commons) to Social Credit, a movement devoted to the heterodox monetary doctrines of Major C.H. Douglas (Ascah 1999). Keynesian macroeconomic policy offered a way to stabilize the economy and avoid depressions without recourse to central planning or inflationary Social Credit (see Brecher 1957, on interwar monetary and fiscal discussions in Canada). William A. Mackintosh of Queen's, nominally only a wartime special assistant to Clifford Clark but de facto head of the Economic Advisory Committee, drafted the federal government's 1945 White Paper on post-war employment policy. The White Paper made a commitment to macroeconomic demand management to maintain full employment that lasted in one form or another for three decades, until in 1975 Bank of Canada Governor, Gerald Bouey, announced the bank's conversion to targeting monetary aggregates to control inflation.

Keynesian ideas reached Canada through Keynes's wartime visits to Ottawa en route to and from the United States, and especially through a group of leading civil servants including some of his former students at Cambridge (Granatstein 1982; Owram 1986). A.F. Wynne Plumptre, who had studied with Keynes in the late 1920s, headed the economics division of the Department of External Affairs and then was Assistant Deputy Minister of Finance (1954–65) before returning to the University of Toronto. Robert Bryce, after attending Keynes's lectures for three years while Keynes was writing *The*

General Theory, was secretary to the Economic Advisory Committee during the war, Secretary to the Cabinet and Clerk of the Privy Council (1954–63), and Deputy Minister of Finance (1963–70). Keynesian macroeconomics reached Canadian academic economists through Mabel Timlin's *Keynesian Economics* (1942). Timlin, a secretary at the University of Saskatchewan, began writing that remarkable book as a Ph.D. dissertation for the University of Washington as early as 1935, before the publication of Keynes's *General Theory,* when Benjamin Higgins arrived in Saskatoon with a copy of Robert Bryce's summary of Keynes's lectures, which Bryce had presented to Hayek's seminar at the London School of Economics, where Higgins was studying. Timlin's book, her first publication at the age of 50, led to a distinguished academic career at the University of Saskatchewan, the presidency of the Canadian Political Science Association, the executive committee of the American Economic Association, and being the first woman in the humanities or social sciences elected to the Royal Society of Canada (see Alexander 1995, on the history of women in economics in Canada). After the war, Timlin wrote a series of review articles in the *Canadian Journal ofEconomics and Political Science* on welfare economics and the applicability of general equilibrium methods to public policy analysis, helping introduce Canadian economists to advances in economic theory elsewhere.

Mabel Timlin was also an early academic critic of the Bank of Canada for permitting inflation during the Korean War by failing to pursue Keynesian stabilization policy. A few years later, many Canadian economists denounced the Bank of Canada Governor, James Coyne, for being more concerned about inflation than with expansionary Keynesian policy to end a recession (Gordon 1961). Economists at the University of Western Ontario, notably David Laidler, Michael Parkin, and Thomas Courchene, later brought to Canada monetarist arguments that the Bank of Canada should adopt a monetary policy rule designed to combat inflation rather than pursuing Keynesian discretionary stabilization policy (Courchene 1975–80).

After the Second World War

The Canadian economics profession expanded along with the great expansion of Canadian universities that began in the 1960s and also with the growing employment of economists in the business community (Parish 1997). Along with the growth of numbers came specialization, first between the different Canadian social sciences (previously sharing departments, conferences and a journal), then between fields within economics. Canadian economics became increasingly theoretical and econometric, and decreasingly historical, in line with changes elsewhere, especially in the United States. Since the rise of academic economics in Canada, Canadian economists had studied in the United States (for example, Innis had taken his Ph.D. at the University of Chicago, with a thesis on the Canadian Pacific Railway) and taken part in American associations, but increasingly Canadian economics, like the rest of Canadian intellectual life, became more oriented towards the United States than to Britain (except that Quebec academics were very conscious of intellectual developments in France). Post-war Canadian economists made noteworthy contributions to economics, particularly the economics of natural resources (Gordon 1954; Scott 1955; Easterbrook 1959; George 1989) and international economics (for example, the effects of trade liberalization), but while Canada's position as a resource-based, small open economy guided the choice of topics, the analytical approaches taken were shared with the international community of economists. Many outstanding economics graduates of Canadian universities pursued careers outside the country, mostly in the United States, but among these, Jacob Viner, John Kenneth Galbraith, Harry Johnson, and Robert Mundell retained close ties to Canada, paid attention to Canada's distinctive economic experience (very large capital inflows relative to GDP before 1914, a floating exchange rate from 1950 to 1962), and took part both in Canadian policy debates and in influencing the development of the Canadian economics profession (for example, Viner 1924; Johnson 1963, 1968).

See Also

- ▶ Galbraith, John Kenneth (1908–2006)
- ▶ Historical Economics, British
- ▶ Innis, Harold Adams (1894–1952)
- ▶ Rae, John (1845–1915)
- ▶ Viner, Jacob (1892–1970)

Bibliography

Alexander, J.A. 1995. Our ancestors in their successive generations. *Canadian Journal of Economics* 28: 205–224.

Angers, F.-A. 1961. Naissance de la pensée économique au Canada français. *Revue de l'histoire de l'Amérique français* 15: 204–229.

Ascah, R.L. 1999. *Politics and public debt: The Dominion, the banks and Alberta's social credit*. Edmonton: University of Alberta Press.

Bladen, V.W. 1978. *Bladen on bladen: Memoirs of a political economist*. Toronto: Scarborough College.

Boucher, P. 1664. *Histoire véritable et naturelle des moeurs et productions de la Nouvelle France.* Paris. Reprinted Boucherville, Québec: Société historique de Boucherville, 1964.

Brecher, I. 1957. *Monetary and fiscal thought in Canada, 1919–1939*. Toronto: University of Toronto Press.

Breckenridge, R.M. 1894. *The Canadian banking system, 1777–1890*. New York: Macmillan.

Brittnell, G.E. 1939. *The wheat economy*. Toronto: University of Toronto Press.

Cherriman, J.B. 1857. Review of A.A. Cournot. Recherches sur les principes mathématiques de la théorie des richesses. *Canadian Journal of Industry, Science and Art* 2: 185–194. Reprinted in Dimand. 1995. Cournot, Bertrand, and Cherriman. *History of Political Economy* 27, 563–578.

Coats, R.H. 1932. Fifty years of statistical progress. In *Fifty years retrospect: Canada 1882–1932*. Toronto: Royal Society of Canada.

Courchene, T. 1975-1980. *Money, inflation and the bank of Canada*. Vol. 2. Montreal: C.D. Howe Research Institute.

Creighton, D. 1937. *The commercial empire of the St. Lawrence 1760–1850*. Toronto: Ryerson.

Dimand, R.W. 1988. An early Canadian contribution to mathematical economics: J.B. Cherriman's 1857 review of Cournot. *Canadian Journal of Economics* 21: 610–616.

Dimand, R.W. 1992. Political protest and political arithmetic on the Niagara frontier: Robert Gourlay's *Statistical Account of Upper Canada*. *Brock Review* 1: 52–63.

Dimand, R.W. 1995. Cournot, Bertrand, and Cherriman. *History of Political Economy* 27: 563–578.

Dimand, R.W. 2005. David Hume on Canadian paper money: A neglected contribution. *Journal of Money, Credit and Banking* 37: 783–787.

Drummond, I.M. 1983. *Political economy at the University of Toronto: A history of the department, 1888–1982*. Toronto: Faculty of Arts and Science, University of Toronto.

Easterbrook, W.T. 1959. Trends in Canadian economic thought. *South Atlantic Quarterly* 58: 91–107.

Falardeau, J.-C. 1944. Problems and first experiments of social research in Quebec. *Canadian Journal of Economics and Political Science* 30: 664–694.

Fowke, V.C. 1946. *Canadian agricultural policy: The historical pattern*. Toronto: University of Toronto Press.

George, P. 1989. Classical writings in Canadian economic history: Natural resources in Canadian economic development. In *Research tools in Canadian studies*, ed. D.C. Poff. Montreal: Association for Canadian Studies.

Goodwin, C.D.W. 1961. *Canadian economic thought: The political economy of a developing nation 1814–1914.* Durham, NC: Duke University Press, and London: Cambridge University Press, for the Duke University Commonwealth Studies Center.

Gordon, H.S. 1954. The economics of a common property resource: The fishery. *Journal of Political Economy* 62: 124–142.

Gordon, H.S. 1961. *The economists versus the bank of Canada*. Toronto: Ryerson.

Gourlay, R. 1822. *Statistical account of upper Canada compiled with a view to a grand system of emigration, in connexion with a reform of the poor laws.* 3 vols. London: Simpkin and Marshall. Reprinted New York: Johnson Reprint Corporation, 1966; abridged with introduction by S.R. Mealing, Toronto: McClelland & Stewart, Carleton Library, 1974.

Granatstein, J.L. 1982. *The Ottawa men: The civil service mandarins 1935–1957*. Toronto: Oxford University Press.

Hamouda, O.F., C. Lee, and D. Mair. 1998. *The economics of John Rae*. London/New York: Routledge.

Helliwell, J. 1993. What have Canadian economists been doing for the past twenty-five years? *Canadian Journal of Economics* 26: 39–54.

Innis, H.A. 1930. *The fur trade in Canada*. New Haven: Yale University Press.

Innis, H.A. 1940. *The cod fisheries*. New Haven: Yale University Press.

Innis, H.A. 1951. *The bias of communication*. Toronto: University of Toronto Press.

Innis, H.A. 1956. In *Essays in Canadian economic history*, ed. M.Q. Innis. Toronto: University of Toronto Press.

Innis, M.Q. 1935. *An economic history of Canada*. 3rd ed, 1954. Toronto: Ryerson.

James, R.W. 1965. *John Rae, political economist: An account of his life and a compilation of his main writings*. Vol. 2. Toronto: University of Toronto Press.

Johnson, H.G. 1963. *The Canadian quandary*. New York: McGraw-Hill.

Johnson, H.G. 1968. Canadian contributions to the discipline of economic science since 1945. *Canadian Journal of Economics* 1: 129–146.
Keyfritz, N., and H.F. Greenway. 1961. Robert Coats and the organization of statistics. *Canadian Journal of Economics and Political Science* 27: 313–322.
Lower, A.R.M. 1943. The development of Canadian economic ideas. In *The Spirit of American Economics*, ed. J.F. Normano, Vol. 1. New York: Day.
Mackintosh, W.A. 1923. Economic factors in Canadian history. *Canadian Historical Review* 4: 12–25.
Mackintosh, W.A. 1939. *The Economic background of dominion-provincial relations.* Ottawa: J.O. Patenaude (King's Printer), Royal Commission on Dominion-Provincial Relations (Rowell-Sirois Commission). Reprinted Toronto: McClelland & Stewart, Carleton Library, 1964.
Mavor, J. 1923. *My windows on the street of the world*. Vol. 2. New York: Dent.
Minville, E. 1979. *L'économie du Québec et la science économique, with a new introduction by F.-A. Angers*. Montreal: Fides.
Montpetit, E. 1939–1942. *La conquetê économique*. Vol. 3. Montréal: Valiquette.
Neill, R. 1972. *A new theory of value: The Canadian economics of H.A. Innis*. Toronto: University of Toronto Press.
Neil, R. 1991. *A history of Canadian economic thought*. London/ New York: Routledge.
Neill, R., and G. Paquet. 1993. L'économie hérétique: Canadian economics before 1967. *Canadian Journal of Economics* 26: 3–13.
Owram, D. 1986. *The government generation: Canadian intellectuals and the state 1900–1945*. Toronto: University of Toronto Press.
Paquet, G., ed. 1989. *La pensée économique au Québec français: Témoinages et perspectives*. Montréal: Association canadienne-française pour l' avancement des sciences, Les Cahiers scientifiques 67.
Parent, E. 1846. *Importance de l'étude de l'économie politique*. Montréal: Revue Canadienne.
Parish, J. 1997. A history of business economics in Canada. *Canadian Business Economics* 5: 1–169.
Parizeau, J. 1968. La recherche en sciences économiques. In *La recherché au Canada français*, ed. L. Beaudoin. Montréal: Presses de l'Université de Montréal.
Pentland, H.C. 1950. The role of capital in Canadian economic development before 1875. *Canadian Journal of Economics and Political Science* 15: 457–474.
Pentland, H.C. 1981. *Labour and Capital in Canada 1650–1860,* ed. with an introduction by P. Phillips. Toronto: James Lorimer & Company.
Rae, J. 1834. *A statement of some new principles of political economy, exposing the fallacies of the system of free trade, and of some other doctrines maintained in the 'Wealth of Nations'.* Boston. Reprint in James. 1965. John Rae, political economist: An account of his life and a compilation of his main writings, Vols. 2. Toronto: University of Toronto Press.
Saunders, S.A. 1939. *The economic history of the maritime provinces.* Ottawa: J.O. Patenaude (King's Printer), Royal Commission on Dominion-Provincial Relations (Rowell-Sirois Commission).
Scott, A.D. 1955. *The economics of natural resources*. Toronto: University of Toronto Press.
Scott Thomas, J. 1889. The conditions of female labour in Ontario. *University of Toronto Studies in Political Science,* series III, ed. W.J. Ashley. Reprint in *The Proper sphere: Woman's place in Canadian Society,* ed. R. Cook and W. Mitchinson. Toronto: Oxford University Press, 1976.
Shortt, A. 1987. *Adam shortt's history of Canadian money and banking: 1600–1880*. Toronto: Canadian Bankers Association.
Shortt, S.E.D. 1976. *The search for an ideal: Six canadian intellectuals and their convictions in an age of transition 1890–1930*. Toronto: University of Toronto Press.
Skelton, O.D. 1911. *Socialism: A critical analysis*. Boston: Houghton, Mifflin.
Skelton, O.D. 1932. Fifty years of political and economic science in Canada. In *Fifty Years in retrospect: Canada*, 1882–1932. Toronto: Royal Society of Canada.
Spafford, S. 2000. *No ordinary academics: A history of the department of economics and political science at the University of Saskatchewan*. Toronto: University of Toronto Press.
Taylor, K.W. 1960. Economic scholarship in Canada. *Canadian Journal of Economics and Political Science* 26: 6–18.
Timlin, M.F. 1942. *Keynesian Economics.* Toronto: University of Toronto Press. Reprint with introduction by L. Tarshis and biographical note by A.E. Safarian, Toronto: McClelland & Stewart, Carleton Library, 1977.
Viner, J. 1924. *Canada s balance of international indebtedness 1900–1913*. Cambridge: Harvard University Press.
Wakefield, E.G. 1968. In *Collected Works*, ed. M.F. Lloyd-Pritchard. Glasgow/London: Collins.

Canard, Nicolas-François (c1750–1833)

R. F. Hébert

Keywords

Canard, N.-F.; Cantillon, R.; Cournot, A. A.; Diffusion theory of taxation; Elasticities of demand and supply; Fuoco, F.; Intrinsic conception of price; Labour theory of value;

Mathematics and economics; Physiocracy; Ricardian theory of land rent; Tax incidence

JEL Classifications
B31

French mathematician and economist, Canard was born in Moulins, near Vichy, around 1750, and died there in 1833. Little is known about his life other than the fact that he taught mathematics at the Ecole Centrale de Moulins. His other interests included economics, jurisprudence and meteorology.

Canard's reputation as an economist rests on his *Principes d'économie politique* (1801), a study of the incidence of taxes, which, however, has drawn more attention for its use of mathematics in economic analysis. Written in the year of Cournot's birth, the *Principes* was honoured by the French Institute, the same body that refused to recognize the later efforts of Cournot and Walras. Cournot (1877, p. i) reviled Canard's work as 'false', even as he admitted that it provided him an important starting point for his own researches. Other harsh critics were Francis Horner, J.B. Say, Joseph Bertrand, W.S. Jevons, and Léon Walras. Despite this rejection by French and English economists, Canard had considerable influence in Italy, where a group of writers, led most conspicuously by Francesco Fuoco, defended his method and adopted some of his ideas. In the present century, Seligman (1927, pp. 159–62) has credited Canard with the diffusion theory of taxation, Schumpeter (1954) has discounted his contribution completely, while Theocharis (1983) has defended him.

The *Principes* was influenced by Cantillon and to a lesser extent by the Physiocrats, whose doctrine Canard sought to refute. Cantillon's influence is obvious in two major areas. First, without using the terms, Canard advanced both an 'intrinsic' and a 'market' conception of price. He held that everything derives its value from the quantity of labour bestowed upon it. Different (unmeasurable) qualities of labour, however, render labour quantity an unsatisfactory measure.

Therefore, one must look to the market to discover the determinants of price. Canard developed an equilibrium theory based on the relative bargaining power of buyer and seller, which he related to need and competition. (Clearly recognizing the forces of monopoly and monopsony, he nevertheless failed to develop a bilateral monopoly model.) Second, Canard revived Cantillon's 'three rents', and wove them into a general equilibrium conception of the economy, which he used to trace the effects of taxation (in the process, adumbrating the Ricardian theory of land rent).

Canard argued that the imposition of a new tax produces disequilibrium and sets in motion certain equilibrating adjustments which take time to work themselves through the economy. Each person who initially pays the new tax will attempt to pass it on to the purchaser of the good, but his success in doing so depends upon the 'forces' encountered; or as we would say today, the tax is shifted in proportion to the elasticities of demand and supply. Canard's maxim that 'every old tax is good, every new tax is bad', must be judged in this context.

Selected Works

1801. *Principes d'économie politique*. Paris: F. Buisson.
1802. *Moyens de perfectionner le jury*. Moulins: P. Vidalin.
1808. *Traité élémentaire du calcul des inéquations*. Paris: Crapelet.
1824. *Eléments de météorologie*. Paris: P. Persan.
1826. *Mémoire sur les causes qui produisent la stagnation et le décroissement du commerce en France, et qui tendent à anéantir l'industrie commerciale*. Paris: Delaunay.

Bibliography

Allix, E. 1920. Un précurseur de l'école mathematique: Nicholas-François Canard. *Revue d'Histoire Economique et Sociale* 13: 38–67.
Baumol, W.J., and S.M. Goldfeld. 1968. *Precursors in mathematical economics: An anthology*. London: London School of Economics and Political Science.
Bousquet, G.H. 1957. N.F. Canard: Précurseur du marginalisme. *Revue d'Economie Politique* 50: 232–235.

Cournot, A.A. 1877. *Revue sommaire des doctrines économiques*. Paris: Hachette.
Schumpeter, J.A. 1954. *History of economic analysis*. London: George Allen & Unwin.
Seligman, E.R.A. 1927. *The shifting and incidence of taxation,* 5th ed. New York: Columbia University Press.
Theocharis, R.D. 1983. *Early developments in mathematical economics,* 2nd ed. London: Macmillan.

Cannan, Edwin (1861–1935)

Murray Milgate

Cannan's name is linked inextricably with two great economic institutions: Adam Smith and the London School of Economics. His edition of the *Wealth of Nations* first appeared in 1904 and remains in print today (1986). Before the publication of the Glasgow edition of Smith's works in 1976, there was nothing that could even lay claim to being its rival. His association with the LSE began as a lecturer in 1895 (the year the School was founded), and continued (in the role of Professor from 1907) until his retirement in 1926.

Cannan was born on 2 February 1861 in Madeira, his mother having gone there on medical advice. Within three weeks of Edwin Cannan's birth his mother had died, and the family returned to Bournemouth where Cannan spent his boyhood. In 1880 he went to Balliol College, Oxford, and took his BA in 1884. He resided in Oxford for the remainder of his life although he was only once formally associated with that city's university when, in 1931, he held the Sidney Ball lectureship. Having a private income from a substantial family fortune, at no time in his life did Cannan have to rely upon securing paid employment for his living. Even after his appointment at the LSE, Cannan was in London on no more than two or three days a week. Cannan was twice President of Section F of the British Association (1902 and 1931), President of the Royal Economic Society (1932–4), and held honorary degrees from Glasgow (LL.D) and Manchester (Litt.D).

To Smith scholarship Cannan bequeathed not only his edition of the *Wealth of Nations* (1904) but also an edition of Smith's Glasgow lectures on jurisprudence (1896). Of the first of these, it is perhaps sufficient to note that subsequent scholarship has modified Cannan's editorial speculations as to its origins in only one major respect – concerning Smith's acquaintance with the work of Turgot – to demonstrate its value. The only other peculiarity of Cannan's commentary concerns his view of the theory of distribution, and it will be necessary to return to this point later. The publication of Smith's Glasgow lectures allowed scholars to observe for the first time just how many of Smith's subsequent views were to be found in his work on economics before his visits to France.

Cannan's original work in the history of economic thought is presented in a number of works, of which two call for separate attention: *A History of Theories of Production and Distribution in English Political Economy 1776–1848* (1893) and A *Review of Economic Theory* (1929). The former is the more carefully considered and better documented of the two, and although it would be difficult to agree with Hugh Dalton who in 1927 claimed that 'no one need ever do this particular piece of work again' (in Gregory and Dalton 1927, p. 11), it is nevertheless the case that both books can be consulted with advantage even by modern students.

It seems that Cannan worked full-time on *Theories of Production and Distribution* from 1890 onwards. In the process of preparing the manuscript, he accumulated a personal library rich in materials form the 18th and 19th century (a library which was subsequently to contain, among other things, a collection of tracts by those Cannan called 'currency cranks' and all editions of Smith's *Wealth of Nations* down to 1900). Many of Cannan's original, if somewhat singular, views gain expression therein. There are two that warrant mention here: the claim that a theory of distribution properly understood requires an explanation of the *shares* of wages, profits and rent in total production (and not an explanation of their respective *rates*, which he calls pseudo-distribution), and the implied definition of

'classical economics' as the period between (and including) Smith's *Wealth of Nations* and the first edition of John Stuart Mill's *Principles* in 1848.

The former opinion is re-iterated in his introduction to the *Wealth of Nations* where it is argued that 'the theory of distribution . . . is no essential part of the work and could easily be excised by deleting a few paragraphs in Book I, chapter vi, and a few lines elsewhere' (1904, p. xxxix). On Cannan's reading, Smith was sidetracked from what should have been his real target (a theory of distribution proper) into a discussion of distribution 'as a mere appendage or corollary of his doctrine of prices' (1893, p. 186), so that 'though Adam Smith had declared that the whole of annual produce is distributed into wages, profit, and rent, obviously meaning thereby total wages, profits, and rent, the last four chapters of Book I of the *Wealth of Nations* deal with wages per head, profits per cent, and rent per acre' (p. 231). Quite how Cannan felt that one might go about explaining his 'distribution properly understood' without a theory about the rates of wages, profits and rent, is impossible to determine from his extant writings. Indeed, his tenacious adherence to this peculiar conception introduces what is perhaps the only real blemish into his editorial introduction of the *Wealth of Nations*. As Higgs observed in his review of that edition in the *Economic Journal* for 1904, Cannan had not so ruthlessly abstained from introducing his own opinions about economic theory as might have been hoped.

Though Cannan does not use the epithet 'classical' to describe either the economics or the economists with which he deals in *Theories of Production and Distribution*, his implied definition of that school in terms of the work on economics in the years between 1776 and 1848 runs counter to the views of those historians of economic thought who prefer to construct a definition of classical economics in terms of some shared set of analytical precepts (a procedure which does not, of course, require that all classical economists reached the same conclusions). It should be noted, however, that Cannan did not subscribe to the view then beginning to emerge that there was a fundamental continuity in the history of economics from 1776 down to the present day. Indeed, like most historians of thought at the time, he was highly critical of Marshall's attempt to establish such continuity in Appendix I of his *Principles* which discusses Ricardo; such views were 'in defiance of all evidence' (1929, p. 177n) as far as Cannan was concerned. There will be cause to return to Cannan's reaction to Marshall later.

The *Review of Economic Theory* (1929) was based on Cannan's LSE lectures to second- and third-year undergraduates (see 1929, p. v). It is an interesting book perhaps more because of what is not said in it rather than what is. It contains no formal presentation of the formula for the elasticity of demand, the treatment of the theory of marginal utility is exceedingly brief, there is no discussion of equilibrium and no reference to the work of Cournot, Pareto, Edgeworth or Wicksell. These latter omissions are striking lacunae – the more so for a book written in the late 1920s. They take on even more significance when it is remembered that this book was an explicit attempt to supplement *Theories of Production and Distribution* with a consideration of work which followed it.

Indeed, it seems that Cannan was no great admirer of the mathematical school, and his opinion of Marshall was to say the least somewhat ambivalent (see, for example, Robbins 1935, p. 396). On this latter point, one may take as an indication his article in *Economica* for 1924 which expresses no great admiration for the quintessentially Marshallian concept of consumer's surplus: it is a method which involves, not a single hypothesis, but an indefinite number of different hypotheses, each of which is inconsistent with all the others as well as with the actual facts . . . inconsistent hypotheses which no one would ever have thought of it if it had not been suggested by the 'space' which happens to be included under the curve of a demand schedule (pp. 23–4). Furthermore, in *An Economist's Protest* (1927) Cannan imagines Adam Smith to comment as follows on 'modern' economics:

> The very ingenious speculations of Mr Jevons, Mr Marshall, Mr Edgeworth and others, . . . have introduced a sort of algebra or geometry into the science

> .... The followers of that system are very numerous; and as men are fond of appearing to understand what surpasses the comprehension of ordinary people, the cypher, as it may be called, in which they have concealed, rather than exposed, their doctrine, [they have] perhaps contributed not a little to increases the number of its admirers (p. 334).

It is doubtful whether the designers of the main doors of the new post-Cannan LSE building understood the irony of their decision to inscribe upon them the now familiar Marshallian demand-and-supply cross diagram.

No account of these two books should omit to mention the ample evidence they provide of Cannan's almost obsessive concern with the etymology of the terms used by economists. Sometimes this propensity led to interesting points, on other occasions it degenerated into farce.

Another of Cannan's contributions to the history of economic thought which may be singled out is his reprint of the Bullion Report (1810), which he published under the title *The Paper Pound* in 1919. The text of the Report runs to 72 pages, Cannan's introduction to it occupies 49 pages. It is of interest not only as an account of the debates which led up to the resumption of specie payments in England with the Act of the Elder Peel in 1819, but also as an indication of the position Cannan was to adopt in the monetary debates of the 1920s and 1930s.

This position was to lead him into head-on collision with the views of Keynes on the question of the advisability of Britain's return to the gold standard after the First World War at the pre-war parity, and his adherence to it, in fact, helps to explain why *The Times* obituary for Cannan was headed 'An Orthodox Economist'. Moreover, the timing of its publication, as Cannan himself observes (1919b, p. xxxix), brings out very clearly the parallels between these two episodes in British monetary history.

Put bluntly, Cannan was probably one of the most strident advocates of the old-fashioned quantity theory around at the time, his solution to inflation being captured in his more than half-serious motto: 'Burn your paper money, and go on burning it till it will buy as much gold as it used to do' (1919b, p. xli). The experience of the policies adopted to deal with the inflation of the post-Napoleonic period were confirmation of the soundness of the return to gold after World War I. Cannan had no sympathy for the idea (still perfectly admissable under the quantity theory) of stabilizing the domestic price level through the management of the domestic supply of money, instead of fixing the exchange rate as the return to gold required. Indeed, it is not always clear from his writings that he understood that the two possibilities were part and parcel of the theory he so vigorously defended.

Nor, it seems, was Cannan prepared to admit the seriousness of the short-run consequences of a policy of deflation on the domestic distribution of income, output and employment as a reason for being cautious about the return to gold – a factor which even someone like Pigou (the official adviser to the Cunliffe Committee and therefore no opponent of the return to gold) was more than prepared to take into account. According to Cannan the necessary adjustments 'must be regarded in the same light as those which a spendthrift or a drunkard is rightly exhorted by his friends to face like a man' (1919b, p. 105).

In addition to the works mentioned above (and those listed in the accompanying bibliography), Cannan contributed twenty-five entries to the original edition of this Dictionary, including those on 'capital' and 'profit'. The latter was cited by Friedman and Savage in their celebrated application of utility analysis to risk. Edwin Cannan was, it is said, a keen bicyclist; though in *Who's Who* he listed his recreation as work. He died on 8 April 1935.

Selected Works

1888. *Elementary political economy.* London: Oxford University Press.

1893. *A history of theories of production and distribution in English classical political economy from 1776 to 1848.* London: P.S. King. 2nd edn, 1903; 3rd edn, 1917.

1896a. *History of local rates in England.* London: Longmans, Green & Co. 2nd edn, London: P.S. King, 1912.

1896b. (ed.) A. Smith: *Lectures on justice, police, revenue and arms.* Oxford: Clarendon Press.
1904. (ed.) A. Smith: *An inquiry into the nature and causes of the wealth of nations,* 2 vols. London: Methuen.
1912. *The economic outlook.* London: T. Fisher Unwin.
1914. *Wealth.* London: P.S. King & Son.
1918. *Money: Its connexion with rising and falling prices.* London: P.S. King & Son. 7th edn, 1932.
1919a. *Coal nationalisation.* London: P.S. King & Son.
1919b. *The paper pound: 1797–1821.* London: P.S. King. 2nd edn, 1925.
1924. 'Total utility' and 'consumer's surplus'. *Economica* 4:21–26.
1927. *An economist's protest.* London: P.S. King & Son.
1929. *A review of economic theory.* London: P.S. King & Son.
1931. *Modern currency and the regulation of its value.* London: P.S. King & Son.
1933. *Economic scares.* London: P.S. King & Son.

Bibliography

Bowley, A.L. 1935. Obituary: Edwin Cannan. *Economic Journal* 45: 385–392.
Gregory, T.E., and H. Dalton (eds.). 1927. *London essays in economics*. London: George Routledge & Sons.
Robbins, L. 1935. A student's recollections of Edwin Cannan. *Economic Journal* 45: 393–398.

Cantillon, Philip (fl. 1725–1759)

Henry Higgs

Author of '*The Analysis of Trade, Commerce, Coin, Bullion, Banks, and Foreign Exchanges:* . . . Taken chiefly from a Manuscript of a very ingenious Gentleman deceas'd, and adapted to the present situation of our Trade and Commerce. By Philip Cantillon, late of the City of London, Merchant, London, 1759'. This Philip was the eldest son of James Cantillon of the city of Limerick, who was first cousin of Richard Cantillon, author of the *Essai sur la Nature du Commerce*. Philip carried on a banking business with David Cantillon at Warnford Court, Throgmorton Street, London, at least as early as 1725. In 1738 he was a director of the Royal Exchange Assurance: in 1742 became bankrupt: in 1747 was trading alone as insurance agent and policy broker: in 1753 was partner with one Thomas Mannock in the same business: and in 1759 had retired. He married, 14 July 1733, Rebecca, daughter of William Newland of Gatton, Surrey, by whom he had two daughters. There is reason to think that he was engaged for a short time at Richard Cantillon's bank in Paris, but that his litigious character made him unamiable and brought about his speedy return. On the death of Richard, Philip intervened in the management of his estate, and thus obtained possession of several papers, including probably the English manuscript of the Essay, which professedly served as the groundwork of the *Analysis of Trade*. He must, however, have mutilated the manuscript almost beyond recognition. Much of the closely packed original is omitted, and much is replaced by vague and general summaries, most unskilfully made, with the result that little indeed of the *Analysis* fairly represents the views of Richard Cantillon. Philip added a preface on the history and importance of commerce, some strictures upon close corporations, new matter on inland and foreign trade, bankers and banks, and exchanges, interspersed with quotations from Hume's *Essays*, and from *The Universal Merchant*, etc., concluding with a criticism of the law relating to bills of exchange.

The book was reviewed in the *Monthly Review* or *Literary Journal* for April 1759, London, vol. xx. 309. Sir James Steuart (*Works*, 1805 edn, iii. 22) says, 'Mr. Cantillon, in his *Analysis of Trade*, which I suppose he understood by practice as well as by theory, has the following passage,' etc.

'A small treatise of Arithmetic,' explaining the foreign exchanges 'vulgarly and decimally'

without 'unintelligible jargon,' was designed by the author of the *Analysis* (p. 85), but does not seem to have ever been published.

Selected Works

1759. *The analysis of trade, commerce, coin, bullion, banks, and foreign exchanges*. London.

Cantillon, Richard (1697–1734)

Vivian Walsh

Keywords

Agricultural surplus; Allocation theory; Cantillon, R; Intrinsic value; Law, J; Luxury; Malthus's theory of population; Market price; Natural price; Produced and non-produced means of production; Quesnay, F; Specie-flow mechanism; Surplus

JEL Classifications

B31

One Richard Cantillon, son of Philip Cantillon of Ballyheigue, County Kerry, was born in Ireland. Joseph Hone argued convincingly that this was the economist, on the ground that this Richard married Mary Ann Mahony, daughter of Lady Clare, and had with her a daughter Henrietta, who married Lord Farnham (after the death of her first husband, the Earl of Stafford). Earlier writers had estimated Cantillon's birth to have been as many as 17 years earlier, but subsequent scholars have tended to accept Hone's evidence; for example, Joseph J. Spengler (1954, p. 283) and Anita Page (1952, p. xxiv).

Richard Cantillon's close association with France has often been noted, but certain facts about his family go far to explaining this connection. An Anglo-Irish county family whose establishment in Ireland was Elizabethan or later would of course be Protestant, and the term 'Anglo-Irish Protestant ascendancy' would then apply strictly. But those families which came to Ireland in Norman times were Catholics, and some of these remained so for hundreds of years, in spite of dungeon, fire and sword (to use an old phrase). They often became Jacobites, and in that case Europe was for them a place of refuge and support. These were the 'Wild Geese', who joined foreign flags after one or other Irish rebellion failed. Often educated in Europe, their ideas were cosmopolitan, their eyes on Paris and on Rome.

The Cantillons were established in Ireland in Norman times and remained Catholics, although not always very good ones. And in later centuries they became, and long remained, devoted to the Stuart cause. Roger Cantillon of Ballyheigue married Elizabeth Stuart in 1556, and his grandson Valentine fought for Charles I at Naseby, while his great-grandson Richard was wounded at the Boyne, went to France with James II and was made a chevalier for his pains. The chevalier, clearly more notable for gallantry than for worldiness, is said to have become banker to the Stuart Pretender in Paris (Spengler 1954, p. 284) and died insolvent, a not unpredictable fate, in 1717. Our Richard appears to have come to the rescue of his uncle's honour, paying off most of the poor old Jacobite soldier's debts, many of which, indeed, were to him. This was not the end of the family's Stuart involvement; a James Cantillon, believed by Hone to be the young future economist's brother, followed King James to France and was decorated for valour, while a nephew, Thomas, mentioned in the economist's will, was with the Irish Brigade at Lauffelt. Migration to France and beyond was in the blood of these wild geese. It should cause no surprise that our Cantillon had houses in seven European cities, or that he lived much in Paris.

He was there, active in banking, between 1716 and 1720. Brilliantly anticipating the fate of John Law's scheme, he was also daring enough to profit immensely by it and, if the sources consulted by W. Stanley Jevons can be believed, 'made a fortune of several millions in a few days, but still, distrusting Law, prudently retired to Holland'

(Jevons 1881, p. 336). He appears again in Paris between 1729 and 1732, and seems to have had to engage in litigation with people who had lost through the collapse of Law's scheme, and blamed Cantillon for his part in this. Henry Higgs, after surveying the evidence, commented that Cantillon appeared 'to have triumphed in the Courts over all his opponents' (Higgs 1931, p. 373). One gets the feeling as one reads of rather ordinary people playing a game for stakes they could not afford with a master they could not match. Bankers fell like autumn leaves in Paris between 1717 and 1720, and as Higgs remarks, 'Their losses were probably very heavy in 1720 and much of them went into Cantillon's pocket' (1931, p. 370).

Back in London in 1734, Cantillon's luck ran out. At the height of his success and his brilliance, he was robbed and murdered, left in the flames of his townhouse in Albermarle Street, Mayfair, during the early morning of 14 May. His precious manuscripts, the Marquis de Mirabeau tells us, perished with him (Higgs 1931, p. 382). Lady Penelope Compton, who lived opposite, tells us that 'it burnt very feirce two houses intirely down before they could get any water' (1931, p. 374). Given this furious blaze, the really remarkable thing to the modern reader is that even despite the primitive state of the forensic science of the day, evidence of foul play was nevertheless found. Higgs, who read the account of the subsequent trial at the Old Bailey, observes that

> it was soon evident that he had been murdered before the house was set on fire. His body was burned to ashes. The Journals for 6 June 1734 say 'Yesterday the refiners finished their search into the ashes of the late Mr Cantillon's house, when no plate, money, or jewels had been found; an undeniable circumstance of a robbery previous to the burning of the house'. (1931, p. 374)

Cantillon's servants were tried for murder, but quickly acquitted. Suspicion then fell on a Frenchman, Joseph Denier, alias Lebane, who, we are told by Higgs, had been Cantillon's cook for 11 years, but apparently had been dismissed a little more than a week before the murder. The French chef, whether in fact guilty or not, fled to Holland and thus evaded arrest.

So it came about that we possess only one work of Cantillon's, and that in what it has been claimed is a rough French translation. Even now its early publishing history is shrouded in mystery. The *Essay on the Nature of Trade in General* (1755) is thought to have been written between 1730 and Cantillon's death, but it was not published in a complete version until 1755, and then in the French translation, claiming on the title page to have been printed in London by Fletcher Gyles, a claim reasonably disputed by Jevons (1881, p. 341). The Marquis de Mirabeau, who revealed that the French translation was in his possession for 16 years, insisted that Cantillon 'never intended that the work should appear in French and only translated it for a friend' (Higgs 1931, p. 383).

Yet, as we have seen, there would be nothing odd in someone of Cantillon's family background and personal habits *writing* a book in French and publishing it in Paris. It would appear, however, that an English original must have existed, and had been in the hands of Malachy Postlethwayt, since the latter incorporated large parts of Cantillon's *Essay* in publications beginning in 1749. The first complete English translation from the French text, which was printed alongside it, was that of Higgs in 1931. Higgs, incidentally, collated his English translation with parallel passages from Postlethwayt. In addition we now have the scholarly French edition, edited by Alfred Sauvy (1952) with a number of studies and commentaries.

Since the 'discovery' of Cantillon by the English-speaking world following Jevons's enthusiastic article (1881), no less than justice has been done to the merits of the *Essay* on those topics treated by Cantillon whose significance can be expressed satisfactorily in broadly neoclassical terms. Over these topics we may pass quickly. Jevons himself noted that Cantillon had presented a treatment of currency, foreign exchanges, banking and credit which, judged against the work of its period, he felt to be 'almost beyond praise' (Jevons 1881, p. 342). This enthusiasm has proved infectious, and we find Joseph Spengler, 73 years later, writing that Hume, assuming he knew Cantillon's work, missed 'the

import of Cantillon's brilliant analysis (which compares favourably with Keynes's) of the response of the price structure to changes in the quantity of money' (Spengler 1954, p. 283). Spengler was not quite as impressed by Cantillon's treatment of the international specie flow mechanism, but Joseph A. Schumpeter found it a brilliant performance and insisted that 'the automatic mechanism that distributes the monetary metals internationally is ... almost faultlessly described' (1954, p. 223).

It was likewise recognized as early as Jevons that Cantillon had set out the leading ideas of Adam Smith's 'important doctrine concerning wages in different employments' (Jevons 1881, p. 343), and that the *Essay* contained what Jevons (somewhat exaggeratedly) called 'an almost complete anticipation of the Malthusian theory of population' (p. 347). Jevons, with remarkable objectivity considering his own views on the formation of value, also singled out Cantillon's treatment of 'the whole doctrine of market value as contrasted to cost value' (1881, p. 345). It was also customarily recognized by neoclassical scholars later than Jevons that Cantillon made important contributions to the founding of allocation theory.

To intellectual historians approaching the *Essay* in terms of the neo-Walrasian class of models for general equilibrium theory, it became natural to construe Cantillon's land and labour as given resources. In the *Essay,* however, while land is a given non-produced input, labour is a *produced commodity* available in return for subsistence. A reproduction structure thus exists, and surplus may be defined. Cantillon is largely concerned with the allocation of surplus output. This was understood by the first classical theorist to read Cantillon, François Quesnay. For all his one-sided preoccupation with *agricultural* surplus, Cantillon's French successor picked up the importance of the role of surplus, embodied it in a formal model and passed it on to later classical economists.

From a *modern* classical point of view Cantillon made several important contributions, which are not always stressed by traditional scholars. For one thing, he offered an early analysis of the respective roles of produced and non-produced inputs in a more than minimally viable commodity reproduction structure. Developing Sir William Petty's concept of a 'par' between land and labour, Cantillon investigated the assumptions upon which a reduction of labour to land is legitimate. But, of course, Cantillon was reducing labour to the *produce* of land; that is, to corn. He noted that 'as those who labour must subsist on the produce of the Land it seems that some relation might be found between the value of labour and that *of the produce* of the Land' (Cantillon 1755b, p. 31; emphasis added). Cantillon had entered an area which even today bristles with problems, which would nowadays be described as concerning the aggregation of heterogeneous objects. Cantillon was well aware of some of them. He used a concept of subsistence, that of the 'meanest Peasant' (p. 39), as his unit of labour, but he was well aware that this differed all over Europe, and had apparently offered statistical material on this in the lost supplement. It is then necessary to be able to express units of more skilled labour in terms of common labour. He argues that 'it is easily seen that the difference of price paid for daily work is based upon natural and obvious reasons' (p. 23). Even today not much progress has been made on this problem, and highly sophisticated models blithely assume it out of existence by using a single homogeneous labour input. Land is also heterogeneous, as Cantillon was well aware; furthermore, any given kind of land can be used to grow different crops. But the analysis of heterogeneous land in the case of a single crop was not developed until Ricardo's period, and the formal analysis of the case where *different crops* are grown had to wait for Piero Sraffa (1960, pp. 74–8), and more recent work on the relations between produced and non-produced means of production, such as that of Alberto Quadrio Curzio (1980, pp. 218–40).

Leaving aside the difficulties of heterogeneous labour and heterogeneous land with multiple uses, the par is the quantity of corn needed for the subsistence of a labourer and his family during a given period. To get a consistent model, corn must be treated as the only commodity strictly necessary to the reproduction system (the only 'basic'

in the Sraffian sense). Other outputs have to be treated as luxury goods (non-basics), so that one can accommodate the changing modes and fashions of Cantillon's prince and landowners. Cantillon in fact allowed even his meanest peasant a number of commodities: 'the married Labourer will content himself with Bread, Cheese, Vegetables, etc., will rarely eat meat, will drink little wine or beer' (Cantillon 1755b, p. 37).

To accept this and retain the par, only two options seem open. The poor peasant's commodities other than bread (or other things made in the household from corn, labour, and any free ingredients) could be regarded as non-basic. Or one could construct a composite commodity, containing bread, cheese, vegetables, and so on, in fixed proportions, and use this as the unit of measurement for the par. Then, if one is to avoid the problems of different crops, one must assume that any parcel of the uniform land can produce these commodities in the standard proportions. Cantillon stressed how much even peasant consumption varied from country to country in Europe in his day. But it was not absurd to suppose, as he did, that consumption habits were fixed and traditional among the peasants of a *particular* area. None of this is meant to deny the justice of Marian Bowley's claim that 'the "par" between land and labour could only be found under special and unrealistic assumptions' (1973, p. 105).

In a model where corn is the only basic, or where a unit of composite commodity is always consumed in fixed proportions, one can express the surplus as corn output minus necessary corn input (seed, subsistence, feed for animals), or alternatively one can express surplus as net output of the composite commodity. Passages such as the following are then consistent with the measurement of the surplus in terms of com (or units of the composite commodity) as required for the par:

> The Farmers have generally two thirds of the Produce of the Land, one for their costs and the support of their Assistants the other for the Profit of their Undertaking ... The Proprietor has usually one third of the produce of his Land and on this third he maintains all the Mechanicks and others whom he employs in the City as well, frequently, as the Carriers who bring the Produce of the Country to the City. (Cantillon 1755b, pp. 43–5)

Cantillon's treatment of surplus strongly implies that it arises only in agriculture. All those in a state, we are told more than once, subsist at the expense of the proprietors of land. There are isolated passages where he seems to be recognizing that profits (in the sense in which these reflect the existence of surplus) can arise in manufacturing. Perhaps the classic case is the description of the master hatter, who, we are told, besides his upkeep, ought also to find 'a profit like that of the Farmer who has his third part for himself' (1755b, p. 203). Certainly Cantillon believed (unlike the Physiocrats) that farmers kept two-thirds of the total produce, one-third representing their *profit.* But Cantillon used his term 'undertaker' (entrepreneur) to cover chimneysweeps and water-carriers, and Samuel Hollander is probably correct in saying that, in Cantillon, 'profits and wages were said to have a common source in, or to be dependent upon, the property of landowners' (1973, p. 40, n. 48). The concept of surplus throughout *industry*, and the dual concept of a rate of profit tending to equality across all sectors, including industrial sectors, would not be clearly and systematically expressed until the mature work of Adam Smith (see Walsh and Gram 1980, pp. 40–77).

Cantillon, however, did pioneering work in developing the theory of the allocation of surplus. His model is remarkably sophisticated. It is an isolated economy – one might think of it as an island – ruled by a prince or landowner. Cantillon is perfectly clear that the prince's significant freedom of choice concerns only that part of output which constitutes the surplus he receives after providing for necessary inputs. He remarks that the prince, deciding on the use of the estate, 'will necessarily use part of it for corn to feed the Labourers, Mechanicks, and Overseers who work for him, another part to feed the Cattle, Sheep and other Animals' (Cantillon 1755b, p. 59). The consumption pattern of workers is fixed, just like fodder for the animals: 'Labourers and Mechanicks who live from day to day change their mode of living only from necessity' (p. 63).

Cantillon is far from assuming, however, that the composition of *surplus* output is unchanging. Indeed, changes in the allocation of surplus,

dictated by changes in the demands of the prince and any other landowners, are his explanation of deviations of current market prices from natural prices, or intrinsic values. In the original classics, and indeed as late as Alfred Marshall (as Pierangelo Garegnani has noted), natural prices are centres of gravitation towards which market prices *tend* (Garegnani 1976). This idea is clearly present in Cantillon. The prince or landlord, who is assumed to have a third of the produce of each of the farms he owns, and is mainly responsible for luxury consumption, is 'the principal Agent in the changes which may occur in demand' (Cantillon 1755b, p. 63). If a few prosperous farmers engage in some luxury consumption, they will imitate the tastes of the prince. Thus changes in fashion were the leading cause of 'the variations of demand which cause the variations of Market prices' (p. 65). Cantillon is well aware that good or bad harvests, extraordinary consumption resulting from foreign troops, and so on, can disturb the gravitation of market prices towards natural prices, but he eliminates such accidents 'so as not to complicate my subject, considering only a State in its natural and uniform condition' (p. 65). This is precisely the concept of a long-period position common to all the great classical economists.

Even more surprisingly, Cantillon shows that he is quite aware that a planned economy directed by the prince, and a system of prices, can each achieve the identical allocation of surplus output – a result whose formal proof had to wait until the 20th century, and which lay fallow after Cantillon as classical political economy developed in other respects.

Cantillon, of course, was by no means the first to make *some* kind of distinction between market and natural prices. The Schoolmen had distinguished between the price ruling at a given moment on a market and the just price, sometimes relating the latter to costs. But in Cantillon the distinction between market and natural price is an integral part of a whole economic model. The natural price, or intrinsic value of a commodity 'is the measure of the quantity of Land and of Labour entering into its production' (1755b, p. 29). Labour is then reduced, through the par, to subsistence units, which, as we have seen, can either be measured in corn or in quantities of a composite commodity. These intrinsic values are assumed to be invariant (p. 31). Market prices may deviate from intrinsic values following a change in demand, as we have seen, but the actions of profit-maximizing capitalist farmers will then lead to supply changes, initiating the gravitation process. If the farmers 'have too much Wool and too little Corn for the demand, they will not fail to change from year to year the use of the land till they arrive at proportioning their production pretty well to the consumption of Inhabitants' (pp. 61–3).

Notice that since we are considering a change in *demand* for corn and wool, these goods are here being used for *luxury* consumption. Corn can be fed to servants and musicians, and wool makes fine garments. What is more, Cantillon can allow for the existence of a number of agricultural sectors producing *only* luxuries: fine wines, silks, blood horses, and so on. His model clearly implies that there is a tendency towards a long-period position in which capitalist farmers in each of these sectors would receive profits at the uniform rate of one-third of the intrinsic value of their total output. Thus the extraction of surplus, and its reflection in a uniform intersectorial rate of profit, is certainly understood by Cantillon for those sectors where capitalist production relations were firmly established in his period. It remained for Adam Smith to extend this analysis to the newly widespread phenomenon of his time, capitalist production throughout industry.

Selected Works

1755a. *Essai sur la nature du commerce en général.* Traduit de l'Anglois, à Londres, chez Fletcher Gyles, dans Holborn.

1755b. *Essai sur la nature du commerce en général.* Ed. with English trans. and other material by Henry Higgs, London: Macmillan (for the Royal Economic Society), 1931.

1952. *Essai sur la nature du commerce en général.* Ed. Alfred Sauvy, Paris: Institut National d'Etudes Demographiques.

Bibliography

Bowley, M. 1973. *Studies in the history of economic theory before 1870*. London: Macmillan.
Garegnani, P. 1976. On a change in the notion of equilibrium in recent work on value and distribution. In *Essays in modern capital theory*, ed. M. Brown, K. Sato, and P. Zarembka. Amsterdam: North-Holland.
Higgs, H. 1931. Life and work of Richard Cantillon. In Cantillon (1755b).
Hollander, S. 1973. *The economics of Adam Smith*. Toronto: University of Toronto Press.
Hone, J. 1944. Richard Cantillon, economist: Biographical note. *Economic Journal* 54(April): 96–100.
Jevons, W.S. 1881. Richard Cantillon and the nationality of political economy. *Contemporary Review* 39 (January). All citations from Higgs (1931).
Page, A. 1952. La vie et l'oeuvre de Richard Cantillon (1697–1734). In Cantillon (1952).
Quadrio Curzio, A. 1980. Rent, income distribution, and orders of efficiency and rentability. In *Essays on the theory of joint production*, ed. L.L. Pasinetti. New York: Columbia University Press.
Schumpeter, J.A. 1954. *History of economic analysis*. New York: Oxford University Press.
Spengler, J.J. 1954. Richard Cantillon: First of the moderns. *Journal of Political Economy* 62(Pt I), 281–295; Pt II, 406–424.
Sraffa, P. 1960. *Production of commodities by means of commodities: Prelude to a critique of economic theory.* Cambridge: Cambridge University Press.
Walsh, V., and H. Gram. 1980. *Classical and neoclassical theories of general equilibrium, historical origins and mathematical structure*. New York: Oxford University Press.

Bibliographic Addendum

For an extended treatment of Cantillon's work, see A. Brewer, *Richard Cantillon*, London: Routledge, 1992.

Capital as a Factor of Production

K. H. Hennings

The role played by capital in production has frequently been in dispute: 'When economists reach agreement on the theory of capital they will shortly reach agreement on everything else' (Bliss 1975, p. vii). Disagreements are due as much to divergent definitions, or uses, of the term 'capital' as to different views about what should be considered a factor of production. But above all there have been differing views about whether, and in what sense, capital can be said to be productive. In particular, there has been disagreement about whether it can be said that a more capital-intensive production method is more productive than a less capital-intensive one. Preclassical, classical, neoclassical and neo-neoclassical economic theory have given different answers to these questions. These will be considered below, but the discussion will be confined to the role of capital as a factor of production. It should be noted in particular that the problem why capital earns its owner an income depends as much on the social institution of ownership and the institutional organization of production as on the role capital plays in production. It is only the latter, in a sense technical, problem which will be addressed here.

Terminology

Capital goods are produced commodities which are required for production no matter how much or how little they are subject to wear and tear. A stock (at a point of time; see Fisher 1906) of different capital goods is a *capital*; this concept is to be taken in a vector sense. As long as they are required in production, all capital goods can be valued, even when they are not traded on markets, as many of them are. Because of their heterogeneity, different capital goods cannot be aggregated, but their values can. A *capital value* is therefore the sum of the capital values of those capital goods which constitute a capital. Note that this is a book-keeping term, which depends on the valuation of the capital goods involved; the capital value can change although there is no change in the stock of capital goods. The term *money capital* will be used in a similar sense, but with a somewhat different connotation: it denotes the sum of money necessary to buy a specified stock of capital goods. Real counterparts in a scalar sense to a given capital value or money capital can be constructed in principle (Hicks 1974, p. 151), but not in an unambiguous manner.

Production: Basic Notions

Production is the transformation of inputs into outputs. Inputs are those things which need to be increased in order to obtain more output by the same method of production, where the latter is defined as a blueprint which details what inputs are required when and in which proportions to produce a unit bundle of outputs. As there may be more than one method to produce the same unit bundle of outputs, a production process is defined as a particular method of production to produce a particular unit bundle of outputs. A production process always uses inputs in fixed proportions; variable proportions are represented by different production processes. If there exist various different production processes with which the same unit bundle of outputs can be produced, they will differ in the proportions in which they use various inputs; but in general it will not be possible to compare them from a purely technical point of view. Different production processes are comparable only if their costs are computed and related to the value of the outputs obtained. In general, however, any ordering obtained in this way need not be unique: two different production processes may have the same unit costs. Moreover, if the prices of inputs change a given ordering need not be preserved. Such difficulties affect the choice between different production processes; they do not, however, affect the role of capital in production, or its status as a factor of production.

Production typically is roundabout, i.e. proceeds in stages: what is produced as output in one production process is used as an input (alongside others) in another. If all these intermediate products (outputs which are used as inputs) are specific in the sense that they have only one possible use, all production processes required to produce a particular bundle of outputs can be strung together into a sequence of production processes. Consolidating all stages, one can view the sequence as transforming 'primary' inputs into 'final' outputs. Here primary inputs are these which are not produced within the sequence of production processes, if indeed they can be produced at all; final outputs are those which are not used, or used up, within the sequence.

Not all intermediate products are specific in the sense that they have only one possible use. In this case all interlocking sequences can be combined into a production system which again can be viewed as transforming primary inputs into final outputs. Without loss of generality one can assume that such a production system comprises all production processes in operation in an economy. Consolidating them amounts to adopting a 'black box' view of production. Disregarding the internal structure of the production system and of the production processes which constitute it, one links directly primary inputs to final outputs, and disregards all inputs produced and used, or used up, within the production system. The advantage of this procedure is that it reduces the number of inputs to be considered.

The definition of what is a primary input, or a final output, depends on the level of aggregation as well as the nature of the production processes involved. Production on a barren island will require many inputs as primary ones which are intermediate products in a production system comprising all production processes operating in a continent rich in resources. Similarly the final outputs produced by the island economy's production system may be confined to what are intermediate products in the production system of a continent.

By definition, an increase in output can only be obtained by an increase in inputs in fixed proportions. From this one can infer that all required inputs together are productive, and have a non-negative marginal production. This cannot, however, be inferred for any single input. This can only be done if either there are at least two different production processes for the production of the same unit bundle of outputs because then it is possible to calculate the marginal net value product of an input (Bliss 1975, ch. 5); or if there are alternative uses for all inputs in production processes which produce other unit bundles of outputs (Uzawa 1958). Only when there exists only one production process for a particular unit bundle of outputs and there are no alternative uses for some of the inputs it requires is it impossible to calculate their marginal contribution to the outputs obtained individually; it is of course still possible to calculate their contribution as a group of inputs.

Factors of Production

In modern usage, all primary inputs can be called 'factors of production'. Conventionally, however, primary inputs are considered, following Senior (1836), the services of agents or stocks, and the term 'factor of production' is reserved for the latter. If they are the services of natural agents or human beings, they are called 'original factors of production'; they are called simply 'factors of production' if they also include the services of stocks of durable commodities. Factors of production can therefore be defined as those agents or durable stocks the services of which are primary inputs in production processes.

Factors of production are productive and have a non-negative marginal product if their services are productive and have a non-negative marginal product.

The definition of factors of production just given is reasonably precise as far as natural agents and human beings are concerned. Land and labour have been considered factors of production at least since Petty (1662). Land was often understood, if tacitly, to include all beneficial powers of nature; the term 'natural agents' was introduced by Senior (1836). In preclassical theory durable stocks were called simply 'stocks' (see, e.g., Barbon 1690), but usage of the term was often confined to trade and commerce. When production came to be seen as the dominant economic activity, produced means of production, considered as a factor of production, came to be called 'capital'. This term had been in use for a long time (see Hohoff 1918–19; Salin 1930; Assel 1953), but now acquired a new meaning, thus inviting confusion and controversy. It will be useful, therefore, to trace historically the use made of that term, and the notions attached to it.

Preclassical Theories of Capital and Production

There is very little about production and its relation to capital in economic writings before the mid-18th century. Barbon (1690) provides an early, but singular, instance of an analysis in which a surplus is seen to arise from the use of what he calls a 'stock' (of capital goods) in trade as well as in the production of commodities. In a similar vein, Hume used the term 'stock' somewhat indiscriminately to denote both a store of commodities and a sum of money. But he did distinguish, as had Barbon, between profits from 'stock' and interest on money (1752, p. 313), thus separating the investment of money from the productive use of 'stock', e.g. capital goods, although he is none too clear about the latter.

The Physiocrats were probably the first to develop a clear view of production and the role of capital in it. But they did not use the term 'capital'. Cantillon (1755) strongly emphasized the need for accumulated sums of money required to buy stocks of goods in which to trade, or with which to produce. But he called them 'funds' not 'capital'. Thus he speaks of the farmer who needs to have enough funds (*assed de fond*) to conduct his business. Quesnay used the term 'advances' (*avances*) in a similar way in the sense of money capital. Behind his usage is a clearly drawn picture of agricultural production which uses land and labour to produce output, and needs money capital to finance the lag between the expenditure on inputs and the sale of the output obtained. Probably deliberately, Quesnay eschewed the term 'capital'. Where he used it (1766b), he spoke explicitly of money capital (*capital d'argent*), but conceived it as invested in buildings, implements, stores of grain, cattle, and so on (1766a, pp. 172–3). These, however, he clearly conceived as productive. Moreover, his argument centres on the idea that larger advances would permit more productive production methods to be used (see Eltis 1984, chs. 1 and 2).

Turgot (1770) was the first to develop a specific theory of capital as a factor in production when, possibly under the influence of Hume's ideas, he generalized Quesnay's theory. Quesnay had shown that advances were necessary for agricultural production. Turgot, in an attempt to develop Quesnay's theory of a society dominated by agriculture into a theory of a commercial society, places commerce and manufacture on an equal footing with agricultural production, and emphasized that advances are required in all branches of

economic activity. Such advances are paid out of capital, which is defined as 'accumulated values' (1770, § LVIII). If account is taken of the various degrees of risks involved, the rates of return on all possible investments are equalized by competition between the owners of the various capitals (Turgot uses the plural, *capitaux*) such that the rate of interest can 'be regarded as a kind of thermometer of the abundance or scarcity of capitals in a Nation, and of the extent of the enterprises of all kinds in which it may engage' (1770, § LXXXIX). At the same time, Turgot argues emphatically that some return on all these kinds of investment is necessary in order to keep production on the same level; if the rate of return were lowered, capitals would be withdrawn, and production could not be kept on the same level as before (1770, § XCVI). Thus to Turgot 'capitals' are money capital. Money capital is required because production is roundabout and thus needs capital goods as well as original factors of production. Like Quesnay, Turgot assumed that larger amounts of money capital make possible higher levels of production. One might be inclined to argue that therefore money capital, i.e. advances, are productive; but although Turgot is not entirely clear on this point it seems that he considered not so much advances as the capital goods which represent them as productive.

The Classical Theory of Capital and Production

The classical view of the role of capital in production was worked out by Adam Smith. He began by emphasizing the division of labour, but then switched to a detailed consideration 'Of the Nature, Accumulation, and Employment of Stock' (1776, book II) in which he effectively adopted the theory put forward by Quesnay and Turgot. His attempts to integrate these two approaches were not entirely successful (Bowley 1976); although the division of labour retained its status as a device which enhances the productivity of labour in classical economic theory, the emphasis was shifted to the accumulation of capital as the prime force making for growth. This was of course linked to the idea that production needs advances, and the proposition that labour was the more productive the larger these advances. Smith also changed the emphasis in another respect: he formally defined 'capital' as that part of a person's stock of commodities which is expected to yield an income. Smith described its function as assisting labour in production: fixed capital (machines, buildings, land improvements, and 'acquired and useful abilities') 'facilitates' labour by increasing its effectiveness; circulating capital (money, raw materials, goods in process and goods in stock) 'abridges' by providing (material) advances.

This distinction is ambiguous, but characteristic for Smith's position. Fixed capital, he argued, yields an income, i.e. is productive, by being used 'without changing masters': while circulating capital needs to be either given up (in trade) or be destroyed (in production) in order to be productive (1776, pp. 279–83). What is considered are capital goods; but only money capital can circulate in the way Smith described their circulation. The two approaches can be reconciled; but the way in which Smith expressed himself invited confusion between money capital on the one hand, and capital in the sense of capital goods on the other. In fact, Smith needed both concepts. James Mill (1821), Rae (1834) and other classical writers often used the term 'instrument' when emphasizing that they meant capital goods, and continued to speak of capital in the sense of money capital. Money capital played indeed an important role in classical economic thought for it permitted classical writers to argue, in a rather loose way, that production methods were the more productive, the more money capital they required. It is for this reason that Hicks (1974) called them 'Fundists'. At the same time, however, they also considered the role of capital goods in production processes (Sraffa 1960), and thus maintained a 'real capital doctrine' (Corry 1962, p. 18).

The view that capital assists labour was attacked by Lauderdale (1804), who pointed out that capital could, and frequently did, supplant labour when circulating capital was substituted for fixed capital. This initiated the debate on the

'machinery question', and confirmed the role of capital as a factor of production: what can supplant a factor of production surely must be considered as belonging to the same species.

Smith had separated a person's stock of commodities into durable consumer goods and capital by requiring that the latter be expected to earn an income. This led to many attempts to show that not only capital goods used in production are expected to yield an income (i.e. Hermann 1832, or Menger 1888). These discussions often confused the role of money capital in investment processes with the role of capital goods in production processes, and contributed to the survival of the concept of money capital as a factor of production referred to above.

The view that production requires advances in the form of capital goods was so dominant that the role of fixed capital was often pushed into the background. Thus Ricardo spoke of production as 'the united application of labour, machinery, and capital' (1817, p. 5), thus equating capital with circulating capital. As Smith had subsumed the consumer goods required for the maintenance of labour under circulating capital stocks, this particular part of the total stock of commodities in an economy acquired, under the name of the 'wages fund', a pivotal role in all discussions of the role of capital in production. Following a precedent set by Smith, the wage fund was seen to be derived from, and increased by, saving, i.e. non-consumption or 'abstinence', as Senior (1836) was to call it. Destined to supply the consumption goods required as advances while production processes continue, the concept was used as a theory of wage determination on the assumption that the wages fund was given at least in the short run and thus determines the wage level when workers compete freely for employment.

In spite of all the attention Smith gave to the accumulation of capital as a factor making for economic growth, he reserved a special role for human labour as the prime factor of production, expecially in those passages in which he set out his conjectural history. This emphasis, which is clearly based on the view that production requires advances, remained a feature of the classical theory of capital, and was a mainstay of the labour theory of value as developed by Ricardo and others. It is symptomatic that from this point of view the use of 'machinery and other fixed and durable capital' was considered no more than an (admittedly considerable) modification of the labour theory of value by Ricardo (1818, p. 30). More radical writers, such as Hodgskin (1827) emphasized the notion of capital goods as 'stored-up' labour (i.e. outputs produced by past labour) that had been worked out by James Mill (1821) and Ricardo (1817) and on its basis denied fixed capital the status as a factor of production.

The special role Smith has reserved for labour did not prevent him from juxtaposing labour, stock and land to parallel wages, profits and rent (1776, p. 69). This juxtaposition was elaborated into a strict parallelism between factors of production and their earnings by Say (1814) which became generally accepted by the middle of the 19th century. Thus when J.S. Mill (1848) summarized the classical theory of capital into his four propositions, he still adhered to the view that production required advances in the form of capital goods. But when he comes to discuss the laws of increase of factors of production, he treats them on an equal footing (though in exactly the order Smith had listed them: and not the land–labour–capital order which Say had made familiar). At the same time, however, Mill often gives the impression that he means money capital when he speaks of 'capital', especially in those passages in which he argues that competition will establish a uniform rate of return on capital because capital will be transferred from one industry to another.

In a similar way Marx (1867) used the term capital to mean both a stock of commodities, and a sum of values. In addition, Marx insisted that capital goods are capital only in a capitalistic society, and thus used the term also to describe a particular organization of production in society.

Finally, the view that production requires advances in the form of capital goods which Smith had expounded, and which most classical writers accepted, was developed by a few of them into a theory which strongly emphasized the time element in production. There are some traces of this in Ricardo (1817), especially in his recognition that all the difficulties he encountered in his

theory of value are due to the temporal aspect of production processes. The view was worked out in detail by Rae (1834), by Longfield (1834), and also by Senior (1836). Their work foreshadows one aspect of the neoclassical theory of capital.

Classical economic theory considered three factors of production: land, labour, and capital. Each had its own dimension: land was a stock, labour a flow, and capital was money capital in the form of a stock of capital goods. In the original conception their standing was not equal: labour worked on land with the help of capital. Hence the capital intensity of production mattered: the more money capital was invested, the more productive was labour in its efforts to work up the bounty of nature into consumable output. These notions were not, however, made very precise: that was left for neoclassical theorists. Thünen's early discussion of the marginal productivity of capital (1850) remained an exception.

The Neoclassical Theory of Capital and Production

Neoclassical economic theory was not a coherent construct: up to the 1930s there were different versions of neoclassical theory as far as the treatment of capital as a factor of production is concerned (Stigler 1941).

Perhaps the most contentious version was the Austrian one as worked out by Böhm-Bawerk (1889). To some extent it had been foreshadowed by Jevons (1871), even though Jevons had little to say about production. But there is a clear picture in Jevons of the necessity of money capital which is 'invested' in the form of advances in time-consuming production processes. What is more, Jevons formulated, very much *ad hoc*, a temporal production function which postulated that there are diminishing marginal returns to the length of investment of such advances: and used it to derive the marginal product of an extension of that length, which clearly is a measure for the capital intensity of production.

Böhm-Bawerk, by contrast, consciously and explicitly developed a theory of production. It very much follows classical lines: production requires time, and hence needs advances in the form of capital goods. Capital goods are seen as produced means of production, and at the same time as stored-up land-and-labour, even though they derive their value not, as the classics had maintained, from the fact that they represent land and labour services spent in the past: but from their prospective usefulness in the production of future output. Nevertheless Böhm-Bawerk emphatically denied that capital goods can be productive, and insisted that only the production processes which they make possible are productive. Although this could have meant that the notion of productiveness was transferred from factors of production to production processes, Böhm-Bawerk did not take this step. He seems to begin by saying that only land and labour should be called productive, and ends by postulating something very much like a productivity of the length of the period of production. As in Jevons (1871), this view is based on a temporal production function in which the degree of roundaboutness of production processes is explicitly taken as a measure for the capital intensity of the production processes in operation. Böhm-Bawerk attempted to overcome in this manner the difficulty of deriving any such measure from diverse sets of capital goods. The roundaboutness of production processes was turned into a variable which was chosen by profit-maximizing entrepreneurs subject to a given amount of money capital.

The relationship of this construction to classical economic thought is obvious. Nevertheless Böhm-Bawerk's attempt to provide a temporal theory of production based on the notion of capital as a derived factor of production, or intermediate good, turned out to be very contentious. The theory of interest which he had been built upon it was turned into what became the standard (neoclassical) theory of interest by Fisher (1907, 1930) – but only after it had been cut loose from its production-theoretic underpinnings: and after Fisher had substituted instead an analysis of investment opportunities based on the concept of money capital. Various attempts to reformulate Böhm-Bawerk's theory of the role of capital in production (Wicksell 1893; Strigl 1934; Hayek

1940) generated much debate, but did not manage to rescue it.

The Austrian theory of capital is much more traditional than other versions of neoclassical theory which gave up the 'advances' view of capital. Thus Wicksteed (1894) placed all factors of production on an equal footing, including all kinds of capital goods, and postulated that 'The Product being a function of the factors of production we have $P = f(a, b, c, \ldots)$' (1894, p. 4) without even mentioning whether production takes time or not. Being considered akin to any other input in this respect, capital goods are of course productive; but nothing can be said about the capital intensity of production. Marshall (1890) argued in a similar way, although he kept to the classical tradition by reserving a place for money capital alongside the capital goods used in production. Taking up a distinction first made, it seems, by Menger (1888, p. 44), Marshall distinguished between capital goods which earn quasi-rents, and money capital which earns interest. In essence this is the distinction between production and investment: capital goods are used in production, and if used productively, earn quasi-rents; money capital is invested, and if invested successfully, earns interest. Clark (1899) equally rejected the advances view of production. In his view, production did not require advances once production processes were properly set up, or synchronized. As in Wicksteed, capital is a factor of production on an equal footing with land or labour. At the same time, Clark separated clearly between material capital goods or produced means of production, on the one hand, and capital as a 'quantum of productive wealth' (1899, p. 119), measured in money, which is invested in capital goods. Although Clark calls this 'a material entity' (1899, p. 119), his 'capital' is money capital, just as it was in Marshall (or Menger for that matter). Knight (1933) continued in this vein, but emphasized money capital, considered as a 'material entity', so much that capital goods were almost lost sight of. As a result, 'capital' came to be seen more and more as a homogeneous mass which was created by saving decisions, which could be invested in one industry and transferred to another, which was productive in the sense that is has a non-negative marginal product if used properly, and which guaranteed higher productivity if employed in larger amounts in relation to other factors of production. Not surprisingly, this conception was attacked by the heirs to the Austrian tradition in capital theory, especially Hayek (1936, 1940), as a 'mythology of capital'. But their own position was so much bound up with the deprecated notion of a period of production that Knight's conception (1933, 1934, 1935, 1936) became the dominant doctrine.

The notion of capital as a 'material entity' was formulated rigorously by Pigou, who provided a sophisticated definition of a capital stock, consisting of heterogeneous capital goods, which 'is capable of maintaining its quantity while altering its form' (1935, p. 239). This was possible only by making some rather strong assumptions on the way the capital stock was maintained. Thus Pigou assumed, among other things, that any item of a constant capital stock that needs to be replaced is replaced by another capital good yielding equal quasi-rents at the time of replacement. Later changes in quasi-rents are disregarded. While such assumptions may be objected to, they do make it possible in principle to give precise meaning to the notion of a capital stock as a changing 'material entity' without aggregating heterogeneous capital goods, i.e. without negating its quality as a vector.

Walras (1874–7) and Pareto (1909) treated capital very much as Wicksteed had done: as yet another factor of production in a production system which was fully synchronized and which was not in need of advances. As they used production functions and thus assumed, as Wicksteed had done, that there always exist many production processes for the production of the same unit bundle of outputs, the productivity of capital goods was no problem for them. But because they espoused the black box view of production they somewhat lost sight of the internal structure of production, and hence of the character of capital goods as produced means of production: capital goods are in their conceptual scheme simply part of the endowment which economic agents use to maximize their satisfaction. Moreover they could not form a notion of the capital intensity

of production as they had no way of aggregating capital goods in an unambiguous manner.

Wicksell, finally, in his later treatment of the matter (1901) attempted to provide a synthesis of neoclassical capital theory by combining the general equilibrium framework of Walras and Pareto with the Austrian view of production as a time-consuming process. This led him to emphasize capital goods and their productivity. But when he came to close his system he took refuge, as Böhm-Bawerk had done, in the idea of a given fund of money capital. The importance of a given fund of money capital which acted as a constraint on entrepreneurial choices between different degrees of roundaboutness of production processes was also emphasized by Schumpeter (1911) and Cassel (1918).

Neoclassical economists have in common that they attempted to formulate a theory of production; but they differed in their conceptions (Hennings 1985). Böhm-Bawerk and those who followed him made an attempt to formulate more precisely what they saw as the gist of the classical theory of the role of capital in production: but their efforts were not generally accepted. All other neoclassical writers except Wicksell jettisoned the advances view of capital, and were in consequence faced with the necessity of formulating a measure for the capital intensity of production if they wished to uphold the proposition that more capital-intensive methods of production were more productive. Wicksteed as well as Walras and Pareto did not do so, and simply refrained from making such statements. Marshall, Clark, and Knight in one way or another attempted to solve the problem by taking refuge in a concept of capital which is in essence a notion of money capital, and which cannot unambiguously serve for that purpose. Only Pigou formulated an unambiguous concept of capital as a changeable 'material entity'.

The Neo-neoclassical Theory of Capital in Production

The neo-neoclassical view of the role of capital in production is based on the work of Viner (1930), Stackelberg (1932), Schneider (1934) and others, who worked out the theory of production as well as a theory of production costs, and the syntheses later provided by Hicks (1939) and Samuelson (1947) of the various neoclassical theories on the basis of the Walras–Pareto theory of general equilibrium (Arrow and Hahn 1971). Originally strongly microeconomic in nature, capital goods held and stage. But as this theory was essentially static, little thought was given to dynamic considerations (Hicks (1939) was the exception), and hence to the problems that arise if concrete capital goods are shifted from one industry to another. Where such problems came up, refuge was taken in the Clark–Knight conception of capital as a fairly homogeneous and amorphous mass which could take on different forms. With the growth of macroeconomic one-sector thinking – Hicks (1932) is one of the earliest examples in this part of economic theory – this conception was more and more resorted to. It received the seal of approval in Samuelson's textbook (1948), and in numerous empirical studies based on the macroeconomic production function first proposed by Cobb and Douglas (1928). It was of course realized that capital consisted of capital goods: but their aggregation into a more or less homogeneous aggregate was considered an index number problem which could be solved in principle as well as in practice. It was against these notions that opposition arose in the 1950s and 1960s.

Recent Debates

As Joan Robinson (1954, 1956) pointed out, the Clark–Knight concept of capital cannot serve in a macroeconomic production function à la Cobb–Douglas because it is essentially a monetary measure. Surprisingly, this contention engendered a major debate in capital theory. Essentially two answers were given to Robinson's objection. On the one hand it was argued that one should search for appropriate indices that can be used to aggregate heterogeneous capital goods into a scalar measure (Champernowne 1954).

This created a specialist literature on aggregation problems which demonstrates that in general

conditions for consistent aggregation are rather restrictive, although in many cases appropriate indices exist (Green 1964). On the other hand, it was argued that macroeconomic analyses should be abandoned in favour of microeconomic ones if heterogeneity (which after all exists in land and labour as well as in capital goods) is the issue (Swan 1962).

In the course of the debates referred to above it was demonstrated that the value paradoxes Joan Robinson had pointed out may invalidate the idea that different production processes can be brought into a continuous ordering which corresponds to their respective capital intensities. While this point was eventually accepted, its importance is still under dispute (see Harcourt (1972) and Blaug (1974) for summaries and evaluations from divergent points of view). To some, such demonstrations completely invalidate neoclassical and in particular neo-neoclassical economic theory, because both are considered to be founded on the idea that marginal products of factors of production need to be calculated on the basis of technical data alone. Others accept such demonstrations as exceptions to a general rule. What is sometimes lost sight of in these assessments is the fact that reswitching of production processes, capital revaluations, Wicksell effects, *et hoc genus omne* do not invalidate all propositions in capital theory (whether neoclassical or not). One can well do without capital in the sense of capital value (i.e. as a scalar magnitude) for some purposes (see, e.g. Nuti 1970). Moreover, it should be appreciated that Robinson's objections do not apply to Pigou's notion of capital as a changeable 'material entity' even though it is not at all obvious that such a concept would serve well in a macroeconomic production function.

Another attack on neoclassical capital theory was made by Garegnani (1960, 1970, 1976). The gist of his argument seems to be that the Walrasian model of general equilibrium, if properly extended to include the production of capital goods, cannot generate equilibrium as well as a unique rate of return on all capital goods for all possible initial endowments. As Garegnani has not specified the dynamic adjustment processes he envisages, his claim is difficult to adjudicate. Nor is it clear in what respect, if any, it invalidates received notions of the role of capital in production processes. Recent debates (Hahn 1982; Duménil and Lévy 1985) have not thrown much light on these issues.

Conclusion

Capital always consists of heterogeneous capital goods; indeed it is useful precisely because goods are heterogeneous and specific in the sense that they cannot be used for all purposes. Attempts to represent them by some kind of aggregate are useful only if they preserve this aspect of capital goods. In classical economic theory the notion of advances was used as such an aggregate, although in a rather loose fashion, with an awareness of the heterogeneity of the capital goods that assisted labour in time-consuming production. Austrian neoclassical economic theory attempted unsuccessfully to make this notion more precise in the form of a temporal theory of production. Non-Austrian neoclassical and neo-neoclassical economic theory sacrificed the heterogeneity of capital goods together with the time element in production, and developed an atemporal theory of production on the basis of a concept of capital value, or money capital. Yet, as Wicksell pointed out (1901, p. 149), the valuation of capital goods in terms of prospective output is a 'theoretical anomaly'; it is nevertheless appropriate in view of their character as produced means of production. It is not surprising, therefore, that anomalies result when such concepts are used. The alternative is obviously to analyse the role of capital goods in a framework which admits their heterogeneity and permits them to be used for different purposes, i.e. in a general equilibrium framework. Such analyses have so far been mainly confined to stationary states. Some of the essential characteristics of capital goods, however, such as their specificity, are of importance only in non-stationary states. Much remains to be done, therefore, before the role of capital and of capital goods as factors of production can be said to be completely elucidated.

See Also

► Roundabout Methods of Production

Bibliography

Arrow, K.J., and F.H. Hahn. 1971. *General competitive analysis*. San Francisco: Oliver & Boyd.

Assel, H.G. 1953. Der Kapitalbegriff und die Kapitallehre bis zum Beginn der Neuzeit. *Wirtschaft und Gesellung. Festschrift für Hans Proesler zu seinem 65. Geburtstag*. Erlangen.

Barbon, N. 1690. *A discourse on trade*. By N.B.M.D. London: Milbourn for the Author.

Blaug, M. 1974. *The Cambridge revolution: Success or failure?* London: Institute of Economic Affairs.

Bliss, C.J. 1975. *Capital theory and the distribution of income*. Amsterdam and Oxford: North-Holland Publishing Company. Innsbruck: Wagner. Trans. as *The positive theory of capital*, London: Macmillan, 1891.

Bowley, M. 1975. Some aspects of the treatment of capital in 'The Wealth of Nations'. In *Essays on Adam Smith*, ed. A.S. Skinner and T. Wilson. Oxford: Clarendon Press.

Cantillon, R. 1755. *Essai sur la nature du commerce en général*. Londres: Gyles. Edited with an English translation by Henry Higgs, London: Cass, 1959.

Cassel, G. 1918. *Theoretische Sozialökonomie*. Leipzig: Deichert. Trans. as *Theory of social economy*, London: Fischer Unwin, 1923.

Champernowne, D.G. 1954. The production function and the theory of capital: A comment. *Review of Economic Studies* 21: 112–135.

Clark, J.B. 1899. *The distribution of wealth*. New York: Macmillan.

Cobb, C.W., and P.H. Douglas. 1928. A theory of production. *American Economic Review* 18: 139–165.

Corry, B.A. 1962. *Money, saving and investment in English economics 1800–1850*. London: Macmillan.

Duménil, G., and D. Lévy. 1985. The classicals and the neoclassicals: A rejoinder to Frank Hahn. *Cambridge Journal of Economics* 9: 327–345.

Eltis, W. 1984. *The classical theory of economic growth*. London: Macmillan.

Fisher, I. 1906. *The nature of capital and income*. New York: Macmillan.

Fisher, I. 1907. *The rate of interest*. New York: Macmillan.

Fisher, I. 1930. *The theory of interest*. New York: Macmillan.

Garegnani, P. 1960. *Il capitale nelle teorie della distribuzione*. Milano: Guiffrè. Trans. as *Le capital dans les théories de la répartition*, Paris: Presses Universitaires de Grenoble et François Maspero, 1980.

Garegnani, P. 1970. Heterogeneous capital, the production function and the theory of distribution. *Review of Economic Studies* 37: 407–436.

Garegnani, P. 1976. On a change in the notion of equilibrium in recent work on value and distribution. In *Essays in modern capital theory*, ed. M. Brown, K. Sato, and P. Zarembka. Amsterdam: North-Holland.

Green, H.A.J. 1964. *Aggregation in economic analysis*. Princeton: Princeton University Press.

Hahn, F.H. 1982. The neo-Ricardians. *Cambridge Journal of Economics* 6: 353–374.

Harcourt, G.C. 1972. *Some Cambridge controversies in the theory of capital*. Cambridge: Cambridge University Press.

Hennings, K.H. 1985. The exchange paradigm and the theory of production and distribution. In *Foundations of economics*, ed. M. Baranzini and R. Scazzieri. Oxford: Blackwell.

Hermann, F.B.W. 1832. *Staatswirthschaftliche Untersuchungen*. Munich: Weber.

Hicks, J.R. 1932. *The theory of wages*, 1963. London: Macmillan.

Hicks, J.R. 1939. *Value and capital*. Oxford: Clarendon Press.

Hicks, J.R. 1974. Capital controversies, ancient and modern. *American Economic Review* 64, May, 307–16. Reprinted in J.R. Hicks, *Economic Perspectives*, Oxford: Clarendon Press, 1977.

Hodgskin, T. 1827. *Popular political economy*. London: Tait and Wait.

Hohoff, W. 1819–1819. Zur Geschichte des Wortes und Begriffes 'Kapital'. *Vierteljahrshefte für Sozial-und Wirtschaftsgeschichte* 14, 554–74 and 15, 281–310.

Hume, D. 1752. Of interest. *Political discourses*. Edinburgh: Fleming. Reprinted in D. Hume, *Essays moral, political and literary*, Oxford: Oxford University Press, 1963.

Jevons, W.S. 1871. *The theory of political economy*. London: Macmillan.

Knight, F.H. 1933. Capitalistic production, time and the rate of return. In *Economic essays in honour of Gustav Cassel*, London: George Allen & Unwin.

Knight, F.H. 1934. Capital, time, and the interest rate. *Economica*, NS 1, August, 257–86.

Knight, F.H. 1935. Professor Hayek and the theory of investment. *Economic Journal* 45: 75–94.

Knight, F.H. 1936. The quantity of capital and the rate of interest. *Journal of Political Economy* 44, August, 433–63, and October, 612–42.

Lauderdale, J. Maitland, 8th Earl of. 1804. *An inquiry into the nature and origin of public wealth*. Edinburgh: Constable.

Longfield, M. 1834. *Lectures on political economy*. Dublin: Milliken.

Marshall, A. 1890. *Principles of economics*. London: Macmillan.

Marx, K. 1867–1894. *Das Kapital*. 3 vols, Hamburg: Meisner. Trans. as *Capital*, 3 vols, Chicago: Kerr, 1906–9.

Menger, C. 1888. Zur Theorie des Kapitals. *Jahrbücher für Nationalökonomie und Statistik, NF* 17(51): 1–49.

Mill, J. 1821. *Elements of political economy*. London: Baldwin, Cradock and Joy.

Mill, J.S. 1848. *Principles of political economy*. London: Longmans, Green.
Nuti, D.M. 1970. Capitalism, socialism, and steady growth. *Economic Journal* 80: 32–57.
Pareto, V. 1909. *Manuel d'conomie politique*. Paris: Giard & Brière. Trans. as *Manual of political economy*, London: Macmillan, 1972.
Petty, W. 1662. *A treatise of taxes and contributions*. London: Brooke. Reprinted in *The economic writings of Sir William Petty*, ed. C.H. Hull, Cambridge: Cambridge University Press, 1899, Vol. 1.
Pigou, A.C. 1935. Net income and capital depletion. *Economic Journal* 45: 235–241.
Quesnay, F. 1766a. (Premier) Dialogue entre Mr. H. et Mr. N. *Journal d'Agriculture, du Commerce et des Finances*, June, 61–109. Reprinted in *Physiocrates*, ed. E. Daire, Paris: Guillaumin, 1846.
Quesnay, F. 1766b. Observations sur l'intérêt de l'argent (par M. Niasque). *Journal d'Agriculture, du Commerce et des Finances*, June, 151–71.
Rae, J. 1834. *Statement of some new principles of the subject of political economy*. Boston: Hilliard Gray & Co.
Ricardo, D. 1817. *On the principles of political economy and taxation*. London: Murray. In *The works and correspondence of David Ricardo*, ed. P. Sraffa, Cambridge: Cambridge University Press 1951, vol. 1.
Robinson, J. 1954. The production function and the theory of capital. *Review of economic studies* 21, 81–106. Reprinted in *Collected economic papers of Joan Robinson*, Vol. II, Oxford: Blackwell.
Robinson, J. 1956. *The accumulation of capital*. London: Macmillan.
Salin, E. 1930. Kapitalbegriff und Kapitallehre von der Antike zu den Physiokraten *Vierteljahrsschrift für Sozial- und Wirtschaftsgeschichte* 23, 401–40.
Samuelson, P.A. 1947. *Foundations of economic analysis*, Harvard Economic Studies, vol. 80. Cambridge, MA: Harvard University Press.
Samuelson, P.A. 1948. *Economics: An introductory analysis*. New York: McGraw Hill.
Say, J.B. 1814. *Traité d'économie politique*, 2nd ed. Paris: Deterville.
Schneider, E. 1934. *Theorie der Produktion*. Vienna: Springer.
Schumpeter, J.A. 1911. *Theorie der wirtschaftlichen Entwicklung*. Leipzig: Duncker & Humblot.
Senior, N.W. 1836. *(An outline of the science of) political economy*. London: Griffin.
Smith, A. 1776. An inquiry into the nature and causes of the wealth of nations. London: Strahan and Cadell. Ed. R.H. Campbell and A.S. Skinner with W.B. Todd, The Glasgow Edition of the Works and Correspondence of Adam Smith, Vol. II, Oxford: Clarendon Press, 1976.
Sraffa, P. 1960. *Production of commodities by means of commodities*. Cambridge: Cambridge University Press.
Stigler, G.J. 1941. *Production and distribution theories*. New York: Macmillan.
Swan, T.W. 1956. Economic growth and capital accumulation. *Economic Record* 32: 334–361.
Thünen, J.H. von 1850. *Der isolierte Staat*. Theil II, 1. Abteilung: *Der naturgemässe Arbeitslohn und dessen Verhältnis zum Zinsfuss und zur Landrente*. Rostock: Leopold. Trans. by B.W. Dempsey as *The Frontier Wage*, Chicago: Loyola University Press, 1960.
Turgot, A.R.J. 1770. Réflexions sur la formation et la distribution des richesses. *Ephémérides du Citoyen*, November and December 1769, January 1770. Trans. as 'Reflections on the formation and distribution of wealth' in *Turgot on progress, sociology and economics*, ed. R.L. Meek, Cambridge: Cambridge University Press, 1973.
Uzawa, H. 1958. A note on the Menger–Wieser theory of imputation. *Zeitschrift für Nationalökonomie* 18: 318–334.
Viner, J. 1930. Cost curves and supply curves. *Zeitschrift für Nationalökonomie* 3: 23–46.
von Hayek, F.A. 1936. The mythology of capital. *Quarterly Journal of Economics* 50: 199–228.
von Hayek, F.A. 1940. *The pure theory of capital*. London: Routledge & Kegan Paul.
von Stackelberg, H. 1932. *Grundlagen einer reinen Kostentheorie*. Vienna: Springer.
von Strigl, R. 1934. *Kapital und Produktion*. Vienna: Springer.
Walras, L. 1874–1877. *Eléments d'économie politique pure*. Lausanne: Corbaz. Trans. as *Elements of pure economics*, London: George Allen 1954.
Wicksell, K. 1893. *Über Wert, Kapital und Rente*. Jena: Fischer. Trans. as *Value, capital and rent*, London: George Allen & Unwin, 1854.
Wicksell, K. 1901. *Föreläsningar i Nationalekonomi*, Första Delen: *Teoretisk Nationalekonomi*. Lund: Berlingska Boktryckeriet. Trans. as *Lectures on political economy*, vol. 1: General Theory, London: Routledge, 1934.
Wicksteed, P.H. 1894. *An essay on the co-ordination of the laws of distribution*. London: Macmillan.

Capital as a Social Relation

Anwar Shaikh

Taken by itself, a sharp stone is simply a relic of some ancient and inexorable geologic process. But appropriated as a cutting instrument, it is a tool or, in a somewhat more murderous vein, a weapon. As a stone, it is a natural object. But as a tool or weapon, it is an eminently social object

whose natural form is merely the carrier of the social relations which, so to speak, happen to have seized upon it.

Even any particular social object, such as a tool, can enter into many different sets of social relations. For instance, whenever a loom is used to weave cloth, it is a part of the *means of production* of a cloth-making labour process. However, because any such labour activity is itself part of the social division of labour, its true content can only be grasped by analysing it as part of a greater whole. For instance, the cloth-making process may be part of the collective labour of a family or community, in which the cloth is intended for direct consumption. Alternately, the very same people may end up using the same type of loom, in a capitalist factory in which the whole purpose of the labour process is to produce a profit for the owners. In the case of cloth produced for direct use, it is properties such as quality and durability which directly concern the producers. But in the case of cloth produced in a capitalist factory, the salient property of the cloth is the *profit* it can generate. All other properties are then reduced to mere vehicles for profit, and as we know only too well, the packaging of the product can easily displace its actual usefulness. This at any rate establishes that even two labour processes which are technically identical can nonetheless have substantially different dynamics, precisely because they exist within very different social frameworks.

The above result also applies to the tools of the labour process. For instance, in both communal and capitalist production, the loom serves as means of production in a labour process. But only in the latter case does it also function as *capital*. That is to say, for its capitalist owners, the significance of the loom lies not in its character as means of production, but rather in its role as means towards profit; while for the workers labouring alongside it, the loom functions not as their own instrument but rather as a proper capitalist tool. Indeed, if we look more closely at the capitalist factory, we will see that not only the loom, but also money, yarn, and even the capacity to labour all serve at various points as particular incarnations of the owners' capital. This is because *capital is not a thing*, *but rather a definite set of social relations* which belong to a definite historical period in human development, and which give the things enmeshed within these relations their specific content as social objects. To understand Capital, one must therefore decipher its character as a social relation (Marx 1894, ch. 48; Marx 1867, Appendix, II–III).

C

Capital and Class

Human society is structured by complex networks of social relations within which people exist and reproduce. The reproduction of any given society in turn requires not only the reproduction of its people, but also of the things they need for their existence, and of the social relations which surround both people and things.

The things which people need for their daily existence form the material base of society. Although the specific character of these things, and even of the needs they satisfy, may vary according to time and circumstance, no society can exist for long without them. Moreover, in all but the most primitive of societies, the vast bulk of the necessary social objects must be produced through human labour. Production, and the social allocation of labour upon which it rests, thus emerge as absolutely fundamental aspects of social reproduction. But social labour involves acting on nature while interacting with other people, in-and-through specific social relations. Thus, the labour process ends up as crucial not only in the production of new wealth, but also in the reproduction of the social relations surrounding this production, as well as of any other social relations directly contingent upon them.

The preceding point assumes particular significance in the case of class societies. In effect, a class society is structured in such a way as to enable one set of people to live off the labour of the others. For this to be possible, the subordinate classes must not only be able to produce more than they themselves appropriate, they must also somehow be regularly induced to do so. In other words, they must be made to work longer than that required by their own needs, so that their surplus labour and corresponding surplus product can be

used to support their rulers. Thus, the very existence of a ruling class is predicated on the *exploitation of labour*, and on the reproduction of the social and material conditions of this exploitation. Moreover, since any such process is a fundamentally antagonistic one, all class societies are marked by a simmering hostility between rulers and ruled, punctuated by periods of riots, rebellions, and revolutions. This is why class societies always rely heavily on ideology to motivate and rationalize the fundamental social cleavage upon which they rest, and on force to provide the necessary discipline when all else fails.

Capitalism is no different in this respect. It is a class society, in which the capitalist class exists by virtue of its ownership and control of the vast bulk of the society's means of production. The working class is in turn comprised of those who have been 'freed' of this self-same burden of property in means of production, and who must therefore earn their livelihood by selling their capacity to labour (labour power) to the capitalist class. As Marx so elegantly demonstrates, the *general social condition* for the regular sale of labour power is that the working class as a whole be induced to perform surplus labour, for it is this surplus labour which forms the basis of capitalist profit, and it is this profit which in turn keeps the capitalist class willing and able to re-employ workers. And as capitalism itself makes abundantly clear, the struggle among the classes about the conditions, terms and future of these relations has always been an integral part of its history (Marx 1867, Part II and Appendix).

Capital as Individual Versus Dominant Social Relation

In the preceding section we spoke about already constituted capitalist society. But no social form springs full blown into being. Instead, its constituent elements must either already exist within other societies, albeit in disassociated form, or else they must arise and be nurtured within the structure of its direct predecessor. This distinction between elements and the whole is important because it allows us to differentiate between capital as an individual social relation, and capitalism as a social formation in which capital is the *dominant* social relation.

Capital as an individual social relation is concerned most of all with the making of profit. In its most general form, this means advancing a sum of money M in order to recoup a larger sum of money M'. The general *circuit of capital* is therefore always attended by the two poles M and M', and their span is always the overall measure of its success. Note that money functions here as a means of making money (i.e. as money-capital), rather than merely as a means of purchasing commodities to be consumed (i.e. as money-revenue). Marx draws many significant and powerful implications from the above functional difference between money-capital and money-revenue.

Even within the circuit of capital, there are three distinct routes possible between its two poles. First, money capital M may be advanced as a loan, in return for a subsequent repayment M' which covers both the original advance and an additional sum over and above it. This is the circuit $M–M'$ of financial capital, in which an initial sum of money appears to directly beget a greater sum, through the apparently magical device of interest. Second, money capital M may be utilized to buy commodities C, and these very same commodities may then be resold for more money M'. This is the circuit $M–C–C–M'$ of commercial capital, in which the double appearance of C as an intermediate term signifies that it is the same set of commodities which first exists as the object of purchase of the capitalist, and then later as their object of (re)sale. Here, it is the acumen of the capitalist in 'buying cheap and selling dear' which appears to generate the circuit's profit. Finally, money capital M may be advanced to purchase commodities C comprising means of production (materials, plant and equipment) and labour power, these latter elements set into motion as a production process P, and the resultant product C' then sold for (expanded) money capital M'. This is circuit $M–C\ldots P\ldots C'–M'$ of industrial capital, in which the characteristic intermediate term is that of the production process P. Now, it is the capitalist's ability to keep the productivity of

labour ahead of the real wage which appears as the fount of all profit.

The most prevalent early incarnations of capital are those of usurer's capital $M–M'$ and merchant capital $M–C–C'–M'$. Both of these are virtually as old as money itself, and have existed over the millennia within many different civilizations. However, they almost always appear as parasitic relations, either within a particular host society or between two or more cultures. Often despised and occasionally feared, these individual activities were nonetheless generally tolerated as long as they conformed to the overall structure of the social formation within which they existed. It is only in feudal Europe, particularly in England, that these antediluvian forms of capital fused together with industrial capital to form the entirely new social formation that we call the capitalist mode of production. Only then, on the foundation of surplus labour extracted directly by itself and for itself, do we find capital as the dominant social relation and its individual forms as mere particular moments of the same overall process (Marx 1858, p. 266, 1867, Appendix).

General Laws of Capital

The social dominance of capital gives rise to certain patterns which are characteristic of the capitalist mode of production.

We have already encountered the first of these, which is that the class relation between capital and labour is a fundamentally antagonistic one, marked by an intrinsic struggle over the conditions and terms of the extraction of surplus labour. Though ever present, this antagonism can sometimes erupt with a force and ferocity which can shake the very foundations of the system itself.

Second, capitalism as a form of social organization pits each element against the other in a generalized climate of conflict: capitalist against worker in the labour process, worker against worker in the competition for jobs, capitalist against capitalist in the battle for market position and sales, and nation against nation in the world market. Like the class struggle, these other conflicts also periodically erupt into acute and open combat between the participants, whether it be the battles of strikers against scabs, or capitalists against their rivals, or even of world wars between one set of capitalist nations and another. It is precisely this real conflict which the bourgeois notion of 'perfect competition' is designed to conceal (Shaikh 1982).

Thirdly, the relations among people are mediated by relations among things. This stems from the very nature of capitalist production itself, in which individual labours are undertaken solely with the aim of making a profit on their product. The various individual labours are thus articulated into a social division of labour only under the 'objectified husk' of their products. It is the products which therefore step to the fore, and the producers who follow behind. From this derives the famous Fetishism of Commodity Relations, i.e. exchangeability appears to be a natural property of all objects, rather than a historically specific way of evaluating the social content of the labour which produced them.

The fourth point follows directly from the third. As noted above, under capitalist relations of production individual labour processes are undertaken in the hope of private gain, with no prior consideration of a social division of labour. But any ensemble of such labours can survive only if they happen to collectively reproduce both the material and social basis of their existence: capitalist society, like all society, requires a particular pattern of labour in order to reproduce its general structure. Thus, under capitalist production, the various individual labours end up being *forcibly articulated into a moving social division of labour*, through a process of trial-through-error, of overshooting and undershooting, of discrepancy, disruption and even occasional ruptures in the process of reproduction. This pattern of apparent anarchy regulated by inner laws of motion is the characteristic form of capitalist reproduction. Notice how different this concept is from that of general equilibrium, where the whole process is reduced to one of immediate and perfect stasis.

The fifth point stems from the fact that capitalist production is driven by profit. Each capitalist is compelled to try and widen the gap between the

initial advance M and the final return M'; those who are most successful prosper and grow, those who fall behind soon face the spectre of extinction. Within the labour process, this shows up in the tendency to stretch the length and intensity of the working day to its social limits, while at the same time constantly seeking to reshape the labour process along lines which are ever more 'rational' from the point of view of capital. This compulsion is directly responsible for capitalism's historically revolutionary role in raising the productivity of labour to new heights. And it is the associated capitalist rationality which is most perfectly expressed in the routinization of production, in the reduction of human activities to repetitive and automatic operations, and in the eventual replacement of the now machine-like human labour by actual machines. As Marx notes, the so-called Industrial Revolution is merely the signal, not the cause, of the advent of capitalist relations of production. And whereas earlier the tool was an instrument of labour, now it is the worker who is an instrument of the machine (Marx 1867, Parts III–IV).

The Conception of Capital Within Orthodox Economics

Within orthodox economics, the term 'capital' generally refers to the means of production. Thus capital, along with labour, is said to exist in every society. From this point of view, social forms are to be distinguished from one another by the manner in which they 'bring together' the factors of production, the capital and labour, at their respective disposals. Capitalism is then defined as a system which utilizes the market to accomplish this task, in the context of the private ownership of the means of production (Alchian and Allen 1983, chs 1 and 8).

By treating human labouring activity as a factor of production on a par with raw materials and tools, *hence as a thing*, orthodox economics succeeds in reducing the labour process to a technical relation between so-called inputs and outputs (e.g. a production function). All struggles over the terms and condition of labour thereby disappear from view.

Moreover, once labour is defined as a factor of production, every (able-bodied) individual is an owner of at least one factor. Of course, some may be fortunate enough to also own large quantities of capital. But that is a mere detail of the distribution of 'initial endowments', and on such things orthodox economics remains studiously neutral. What matters instead is that under capitalism the notion that everybody owns a factor of production bespeaks of an inherent equality among individuals. Any reference to the concept of class is therefore blocked from the start.

Next, because labour is merely one of the factors of production which individuals are free to utilize in any manner they choose, this labour-as-thing cannot be said to be exploited. The exploitation of labour thus drops out of sight, to be replaced by the notion of the cooperation of Capital and Labour, each of which contributes its component to the product and receives in turn its commensurate reward (as in marginal productivity theories of distribution). With this, the sanctification of capitalism is complete.

The Historical Limits of Capital as a Social Relation

The last general point has to do with the historical specificity of capitalist production. On the one hand, capitalism is a powerful and highly flexible social structure. It has developed its forces of production to extraordinary heights, and has proved itself capable of dissolving or destroying all previous social forms. Its inherently expansive nature has led to the creation of vast quantities of wealth, and to a dominion which extends all over the globe. But on the other hand, this very same progressive aspect feeds off a dark and enormously destructive side whose nature becomes particularly clear when viewed on a world scale. The capital-labour relation is a profoundly unequal one, and the concentration and centralization of capital which attends capitalist development only deepens the inequality. The competitive struggle of all against all creates an alienated and selfish social character, imprisons each in an atmosphere of suspicion and stress, and heaps its

miseries precisely on those who are in the weakest positions. Finally, as capitalism develops, so too does its level of mechanization, so that it is progressively less able to absorb labour. In the developed capitalist countries, this manifests itself as a growing mass of unemployed people at any given 'natural' rate of unemployment. In the Third World, as the incursion of capitalist relations lays waste to earlier social forms, the mechanized processes which replace them are able to pick up only a fraction of the huge numbers previously 'set free'. Thus the rising productivity of capitalist production is accompanied by a growing pool of redundant labour all across the globe. The presence of starving masses in the Third World, as well as of floating populations of unemployed in the developed capitalist world, are bitter reminders of these inherent tendencies.

The above perspective forcibly reminds us that capitalism is only one particular historical form of social organization, subject to deep contradictions which are inherent in the very structure of its being. Precisely because these contradictions are built-in, any successful struggle against their destructive effects must move beyond reform to the rejection of the structure itself. In the 20th century such efforts have taken a variety of forms, ranging from so-called parliamentary socialism to socialist revolution. Whatever we may think of the strengths and weaknesses of these various fledgling social movements, the general tendency is itself part of an age-old human process. History teaches us that no social form lasts forever, and capital as a social relation is no exception to this rule.

See Also

► Class

References

Alchian, A.A., and W.A. Allen. 1983. *Exchange and production: Competition, coordination, and control*, 3rd ed. Belmont: Wadsworth Publishing Co.

Mandel, E. 1976. *Introduction to vol. I of Capital by K. Marx (1867)*. London: Penguin.

Marx, K. 1858. *Grundrisse.* London: Penguin, 1973.

Marx, K. 1867. *Capital*, vol. I. London: Penguin, 1976.

Marx, K. 1894. *Capital*, vol. III. Introduced by Ernest Mandel. New York: Vintage, 1981.

Rosdolsky, R. 1977. *The making of Marx's capital*. London: Pluto Press.

Shaikh, A. 1982. Neo-ricardian economics: A wealth of algebra, a poverty of theory. *Review of Radical Political Economics* 14(2): 67–83.

Capital Asset Pricing Model

M. J. Brennan

Abstract

Two general approaches to the problem of valuing assets under uncertainty may be distinguished. The first approach relies on arbitrage arguments of one kind or another, while under the second approach equilibrium asset prices are obtained by equating endogenously determined asset demands to asset supplies, which are typically taken as exogenous. The capital asset pricing model (CAPM) is an example of an equilibrium model in which asset prices are related to the exogenous data, the tastes and endowments of investors, although the CAPM is often presented as a relative pricing model.

Keywords

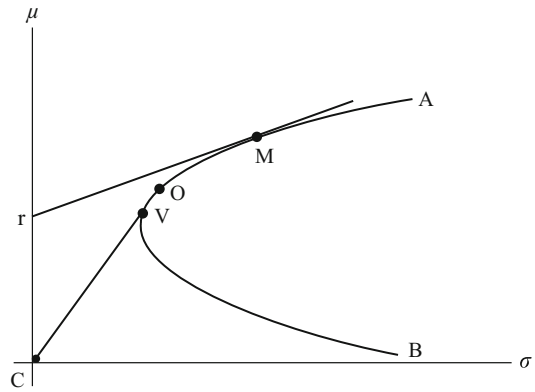
Arbitrage pricing theory; Asset price anomalies; Capital asset pricing model; Consumption capital asset pricing model; International asset pricing model; Intertemporal capital asset pricing model; Mean-variance analysis; Portfolio theory; Pricing kernels; Probability distributions; Relative pricing models; Risk aversion; Separation; Tobin separation theorem; von Neumann and Morgenstern

JEL Classifications

G12

If they are to be of practical use, equilibrium asset pricing models must be parsimonious in their

parameterization of asset demands. To date this parsimony has been achieved only by a choice of assumptions which leads to universal portfolio separation: this is the property that the asset demand vector of every agent can be expressed as a linear combination of a set of basis vectors which may be thought of as portfolios or mutual funds. The distinguishing feature of the set of models which is collectively known as the capital asset pricing model (CAPM) is that each of these basis portfolios can be interpreted as the solution to a particular constrained portfolio variance minimization problem.



Capital Asset Pricing Model, Fig. 1 The efficient frontier and the CAPM

Historical Perspective

The assumption that uncertainty about future asset returns can be described in terms of a probability distribution is at least as old as Irving Fisher (1906), although Hicks (1934b) appears to have been the first to suggest that preferences for investments could be represented as preferences for the moments of the probability distributions of their returns, and to propose that, as a first approximation, preferences could be represented by indifference curves in mean-variance space. Von Neumann and Morgenstern (1947) were the first to place the theory of choice under uncertainty on a rigorous axiomatic basis.

The story of modern portfolio theory really begins, however, with Markowitz (1952, 1958) who assumed explicitly that investor preferences were defined over the mean and variance of the aggregate portfolio return, related these parameters to the portfolio composition and the parameters of the joint distribution of security returns, and for the first time applied the principles of marginal analysis to the choice of optimal portfolios.

Both Markowitz and Tobin (1958) showed that mean-variance preferences can be reconciled with the von Neumann–Morgenstern axioms if the utility function is quadratic in return or wealth. This assumption is objectionable since it implies negative marginal utility at high wealth levels. Tobin also showed, however, that mean-variance preferences could be derived by restricting the probability distributions over which choices are made to a two-parameter family. After some initial confusion it was recognized that, since portfolio returns are weighted sums of security returns, the two-parameter family must be stable under addition, and the only member of the stable class with a finite variance is the normal distribution. Subsequently Merton (1969) and Samuelson (1970) showed that mean-variance analysis is applicable for a broad class of continuous asset price processes if the trading interval is infinitesimal.

The major part of Tobin's analysis deals with the choice between a single risky asset and cash, but he demonstrated that nothing essential is changed if there are many risky assets, for they will always be held in the same proportions and can be treated as a single composite asset. This, the first separation theorem in portfolio theory, is illustrated in Fig. 1, which plots mean returns, μ, against the standard deviation, σ. In this figure the curved locus AMOVB corresponds to the set of portfolios offering the lowest standard deviation for each level of mean return: the positively sloped segment is referred to as the efficient frontier, for points along it offer the highest μ for a given σ. In the absence of any riskless investment opportunities, risk-averse mean-variance investors will select portfolios corresponding to the points at which their indifference curves in (μ, σ) space are tangent to the efficient frontier (Tobin shows that the indifference curves of risk averters will have the requisite curvature). Point C

represents cash which has zero risk and return. By combining cash with the portfolio of risky assets corresponding to the tangency portfolio O, investors are able to attain the (μ, σ) combinations along the line segment CO, and all investors who find it optimal to hold cash will find it optimal to combine their cash with the same risky portfolio O: their portfolio decisions can be *separated* into the choice of the optimal combination of risky asset (O) and the choice of the cash–risky asset ratio.

Six years elapsed before the equilibrium implications of the Tobin separation theorem were exploited by Sharpe (1964) and Lintner (1965). The reason for delay was undoubtedly the boldness of the assumption required for progress, namely, that all investors hold the same beliefs about the joint distribution of security returns. Nevertheless, this assumption of homogeneous beliefs, combined with the further assumption that all investors can borrow as well as lend at the riskless rate, r, leads to the powerful conclusion that all investors hold the same portfolio of risky assets, denoted by M in the figure. Then the only risky assets that will be held by investors in equilibrium are those contained in portfolio M, and M must be the market portfolio of all risky assets in the economy. This identification of the tangency portfolio M with the aggregate market portfolio is the essence of the Sharpe–Lintner CAPM.

The interest of this result derives from the restriction that it imposes on expected asset returns: the excess of μ_j, the expected return on any security j, over the risk-free rate r, must be proportional to the covariance of the security return with the return on the market portfolio, σ_{jM}:

$$\mu_j - r = \theta_M \sigma_{jM} \text{ for all } j \tag{1}$$

where θ_M is a measure of aggregate risk aversion. The intuition behind this important result is that if investors are content to hold portfolio M, the marginal rate of transformation between risk and return obtained by borrowing to invest in a risky security must be the same for all risky securities. Frequently the unknown risk aversion parameter, θ_M, is eliminated and the relative pricing result is obtained:

$$u_j - r = \beta_j(\mu_M - r) \text{ for all } j \tag{2}$$

where μ_M is the expected return on the market portfolio and $\beta_j \equiv \sigma_{jM}/\sigma_{MM}$ is the 'beta' coefficient, which corresponds to the slope of the regression line relating the return on the security to the return on the market portfolio.

During the first half of the 1970s extensive progress was made in relaxing the strong assumptions underlying the original model, and new separation theorems and models were obtained. At the same time, extensive empirical investigations made possible by the development of new stock-price databases found results which were interpreted as favourable to the model. The model also has an influence on practical investment management and corporate finance.

A turning point was reached with the publication of a paper by Roll (1977); this argued that the market portfolio of the theory, which includes all assets, could never be empirically identified, and that therefore the CAPM, which simply asserts the efficiency properties of this portfolio, could never be empirically tested. This argument had substantial influence, and for some time played a major role in shifting attention away from the CAPM to the newly emerging arbitrage pricing theory (APT) of Ross (1976). However, since the early 1990s growing acceptance of the empirical importance of time variation in investment opportunities has led to a resurgence of interest in Merton's (1973) intertemporal version of the CAPM which is formally similar to the APT but is able to provide an economic interpretation of the return factors that are priced in equilibrium.

The CAPM is of great historical significance, not only because it was the first equilibrium model of asset pricing under uncertainty, but also because it showed the importance of portfolio separation for tractable equilibrium models; and, being derivable from assumptions of either quadratic utility or normal distributions, it revealed that the requisite separation properties could be obtained by restrictions either on preferences or on distributions. Cass and Stiglitz (1970) clarified

the rather restrictive assumptions necessary for preference-based separation, and equilibrium models based on this have been constructed, for example, by Rubinstein (1976). Ross (1978) has identified the distributional assumptions required for separation in the absence of restrictions on preferences, and the arbitrage pricing theory is based on a generalization of his separating distributions. Chamberlain (1983) discusses spherical distributions, the subclass of separating distributions for which the expected utility is a function of the portfolio mean and variance. Both preference-based and distribution-based models of capital market equilibrium are lineal descendants of the CAPM.

A pricing kernel is a non-negative weighting function for asset returns under which the expected returns on all assets are equal to the risk-free interest rate; the kernel corresponds roughly to the marginal utility of a representative investor and the existence of a pricing kernel is a necessary and sufficient condition for arbitrage free security markets. Modern treatments of asset pricing such as Cochrane (2005) treat the general problem of asset pricing as that of specifying an appropriate pricing kernel: the CAPM specifies a class of pricing kernels that are linear in the aggregate market return.

An unfortunate consequence of the one-period nature of the CAPM was a concentration of attention on equilibrium rates of return, rather than on prices, which are the fundamental variables of interest. However, Merton (1973) placed the CAPM in an intertemporal context, and his necessary condition for equilibrium rates of return forms one cornerstone (the other being an assumption of rational expectations) for partial differential equations for asset prices which, following Cox et al. (1985), has tended to unify the pricing theories for bond and equity markets.

Formal Models

While a complete asset pricing model endogenizes the riskless interest rate as well as the prices of risky securities, the CAPM adds nothing new to the theory of interest rate determination, and we shall simplify by taking the interest rate and current consumption decisions as given, concentrating our attention on portfolio decisions and the pricing of risky securities.

In considering the various versions of the CAPM we shall pay particular attention to the implied demands of investors. It will be seen that in all cases in which risks are freely traded asset demands exhibit the separation property, and even when there are restrictions on trading as in the Mayers (1972) asset pricing model, an approximate separation property obtains.

The Sharpe–Lintner Model

Consider a setting in which each investor $i(i = 1,\ldots,m)$ is endowed with a fraction $\overline{z}_{ij}$ of security $j(j = 1,\ldots,n)$ and (a) investor utility is defined over the mean and variance of end of period wealth; (b) securities are traded in a competitive market with no taxes or transactions costs; (c) investors share homogeneous beliefs or assessments of the joint distribution of payoffs on the securities; there are no dividends; (d) there is an exogenously determined interest rate $r = R - 1$ at which investors may borrow or lend without default; (e) there are no restrictions on short sales.

Then define:

$\overline{p}_{j1}$ expected end of period value of security j;
$\overline{p}_{j0}$ initial value of security j;
ω_{jk} covariance between end of period value of j and k;
$\overline{W}_i, S_i^2$ expectation and variance of end of period wealth of investor i;
$V_i(\overline{W}_i, S_i^2)$ utility of investor i with
$V_{i1} \equiv \partial V_i / \partial \overline{W}_i > 0, V_{i2} \equiv \partial V_i / \partial S_i^2 < 0.$

The investor's decision problem may be written as

$$\max_{z_{ij}} \; V_i\left(\overline{W}_i, \mathrm{S}_i^2\right) \tag{3}$$

$$\text{s.t. } \overline{W}_i = \sum_j z_{ij}\overline{P}_{j1} - R\sum_j \left(z_{ij} - \overline{z}_{ij}\right)P_{j0} \tag{4}$$

$$S_i^2 = \sum_j \sum_k z_{ij} z_{ik} \omega_{jk}. \tag{5}$$

The first order conditions for an optimum are

$$V_{i1}\left(\overline{P}_{j1} - RP_{j0}\right) + 2V_{i2}\sum_k z_{ik}\omega_{jk} = 0, \quad (j = 1, \ldots, n) \tag{6}$$

and the second conditions are satisfied by virtue of the assumption of risk aversion. Defining $\mathbf{\Omega}^*$ as the variance covariance matrix $[\omega_{jk}]$ and using boldface type to denote vectors, the vector of fractional asset demands may be written

$$\mathrm{z}_i = \theta_1^{-1}\mathbf{\Omega}^{*-1}\left(\overline{\mathbf{P}}_1 - R\mathbf{P}_0\right) \tag{7}$$

where $\theta_i^1 \equiv -V_{i1}/2V_{i2}$ is a measure of the investor's risk tolerance. Equation (7) is a statement of the Tobin separation theorem, that investor demands for risky assets differ only by a scalar multiple.

Market clearing requires that $\sum_i \mathbf{Z}_i = 1$where **1** is a vector of units. Then the equilibrium initial price vector is obtained by summing (7) over i and imposing the market clearing condition:

$$\mathbf{P}_0 = \frac{1}{R}\left\{\overline{\mathbf{P}}_1 - \theta_m \mathbf{\Omega}^* 1\right\} \tag{8}$$

where $\theta_m \equiv \left(\sum_i \theta_i^{-1}\right)^{-1}$. In this form the CAPM expresses equilibrium asset prices in terms of the exogenous variables, the distribution of end of period prices, investor risk aversion parameters and the interest rate, although it should be noted that in general the market risk aversion parameter $\theta_m$will depend upon the endogenously determined distribution of wealth. This formulation corresponds to that of Lintner (1965) and emphasizes the one-period nature of the model and the exogeneity of the end of period prices. However, the CAPM is most often written as a necessary condition for the equilibrium rates of return, although this obscures the distinction between endogenous and exogenous variables.

In what follows we shall work with the rate of return formulation; thus define $x_{ij} \equiv z_{ij}P_{j0}$, the amount invested in security j; $\mu_j \equiv \overline{P}_{j1}/\mathrm{P}_{j0} - 1$, the expected rate of return and$\sigma_{jk} \equiv \omega_{jk}/P_{j0}P_{k0}$, the covariance of the rates of return between securities j and k. Making these substitutions in (4) and (5), the first order conditions (6) become

$$V_{i1}\left(\mu_j - r\right) + 2V_{i2}\sum_k x_{ik}\sigma_{jk} = 0, \quad (j = 1, \ldots, n). \tag{9}$$

Then, defining $\mathbf{\Omega}$ as the variance covariance matrix of rates of return, the vector of asset demands $\mathbf{x}_i$ may be expressed as

$$\mathbf{x}_i = \theta_i^{-1}\mathbf{\Omega}^{-1}(\boldsymbol{\mu} - r1). \tag{10}$$

This is an alternative statement of the Tobin separation theorem and the portfolio $\mathbf{\Omega}^{-1}(\boldsymbol{\mu} - r1)$ corresponds to the point of tangency in Fig. 1. This portfolio itself may be decomposed into the two portfolios $\mathbf{\Omega}^{-1}\boldsymbol{\mu}$ and $\mathbf{\Omega}^{-1}1$. The former is the solution to the problem of finding the minimum variance portfolio of risky assets with a given expected payoff, and the latter is the solution to the problem of finding the global minimum variance portfolio of risky assets; these two portfolios plot at points O and V in the figure. As Merton (1972) has shown, the whole locus may be constructed from just these two portfolios.

Let Vm denote the aggregate market value of all assets in the market portfolio and let $\mathbf{v}_\mathrm{m}$ denote the vector of market proportions. Combining the market clearing condition $\sum_i \mathbf{x}_i = V_m\,\mathbf{v}_\mathrm{m}$ with (10) yields

$$\boldsymbol{\mu} - r1 = \theta_m V_m \mathbf{\Omega}\mathrm{v}_m. \tag{11}$$

This form of the CAPM expresses asset risk premia as proportional to the covariances of their returns with the returns on the market portfolio; this of course is no more than the condition for the market portfolio to correspond to the tangency point in Fig. 1. Equation (11) contains the market risk aversion parameter θ_m. This can be eliminated by pre-multiplying (11) by $\mathbf{v_m}$ and solving for $\theta_m = (\mu_m - r)/\sigma_m^2$, where μm and $\sigma_m^2$are the

C

expected return and variance of return on the market portfolio respectively. Then, substituting for θ_m in (11) we have the equation of the 'security market line':

$$\mu_j - r = \beta_j(\mu_m - r) \tag{12}$$

where $\beta_j \equiv \sigma_{jm}/\sigma_m^2$. In this form the CAPM is a relative pricing model which relates the risk premium on individual securities to the risk premium on the market portfolio. The proportionality factor, β_i, often referred to as the 'beta coefficient', is the coefficient from the regression of $\tilde{R}_j$, the return on security j, on $\tilde{R}_m$, the return on the market portfolio:

$$\tilde{R}_j = \alpha_j + \beta_j \tilde{R}_m + \tilde{e}_j \tag{13}$$

where $\tilde{e}_j$is an orthogonal error term. Taking expectations in the market model Eq. (13), the asset pricing Eq. (12) is seen to imply the restriction $\alpha_j = (1 - \beta_j)r$. This restriction, and the existence of a positive risk premium on the market portfolio, are the major empirical predictions of the Sharpe–Lintner model. They have been the subject of extensive empirical tests.

Taxes and Restrictions on Riskless Transactions

The absence of short sales restrictions is not critical to the Sharpe–Lintner model, since in equilibrium all investors hold the market portfolio, which does not involve short sales. The assumption is critical, however, for all the remaining models we shall consider which involve more than a single basis fund of risky securities.

Thus, following Black (1972) and Brennan (1970), assume that there are no opportunities for riskless borrowing or lending, and that each security pays predetermined dividends which are taxed in the hands of the investor at the rate $t_i (i = 1, \ldots, m)$. Denoting the dividend yield by δ_j, and assuming that investor preferences are defined over the moments of after tax wealth, the first order conditions corresponding to (9) are

$$V_{i1}\left(\mu_j - t_i\delta_j - \lambda_i\right) + 2V_{i2}\sum_k x_{ik}\sigma_{jk} = 0, \quad (j = 1, \ldots, n). \tag{14}$$

where λ_i is the Lagrange multiplier associated with the constraint that all wealth be invested in risky securities. The vector of asset demands may be written as

$$x_i = \theta_i^{-1}\boldsymbol{\Omega}^{-1}\boldsymbol{\mu} - \left(\theta_1^{-1}\lambda_1\right)\boldsymbol{\Omega}^{-1}1 - \left(\theta_1^{-1}\mathbf{t}_1\right)\boldsymbol{\Omega}^{-1}\delta. \tag{15}$$

Note first that if $t_i = 0$ the optimal portfolio for any preferences can be constructed from the two mutual funds $\boldsymbol{\Omega}^{-1}\boldsymbol{\mu}$ and $\boldsymbol{\Omega}^{-1}\mathbf{1}$. Heterogeneous taxation of dividends introduces the third mutual fund, which can be interpreted as the solution to the problem of finding the minimum variance portfolio with a given total dividend. Aggregating the demand vectors, and imposing the market clearing conditions, yields an asset pricing equation which contains three utility dependent parameters, λm, θ_m and t_m, corresponding to the three funds in (15):

$$\boldsymbol{\mu} - \lambda_m 1 = \theta_m V_m \boldsymbol{\Omega} \mathrm{v}_m + t_m \delta \tag{16}$$

t_m, the market tax rate, is a weighted average of the personal tax rates, and λ_m, the market shadow interest rate, is referred to for historical reasons as the zero beta return. When $t_m = 0$, (16) is just the condition for the market portfolio to be the tangency portfolio when the interest rate is λ_m. Thus the Black model, which does not include taxes, differs from the Sharpe–Lintner model only in leaving unspecified the relevant (shadow) riskless interest rate.

Non-marketable Assets

Mayers (1972) has considered the effect of introducing an extreme form of market imperfection, namely, an absolute prohibition on trading certain assets. This is important, for a substantial part of total wealth is not held as part of well-diversified portfolios, on account either of prohibitions on trade (human capital), or of market imperfections

such as transactions costs and information asymmetries. Thus let $\overline{h}_i$ denote the expected payoff on the non-marketable wealth (human capital) of investor i, and let σ^i_{jh} denote the covariance between the return on marketable security j and the human capital of investor i. Then the expression for $\overline{W}_i$ must be increased by $\overline{h}_i$ and the variance of end of period wealth becomes $S_i^2 = \sum_j \sum_k x_{ij} x_{ik} \sigma_{jk} + 2\sum_j x_{ij}\sigma^i_{jh} + \sigma$. The asset demand vector can then be written as

$$\mathbf{x}_i = \theta_i^{-1} \mathbf{\Omega}^{-1}(\boldsymbol{\mu} - r1) - \mathbf{b}_i \tag{17}$$

where $\mathbf{b}_i = \mathbf{\Omega}^{-1}\sigma^i_n$ is the vector of coefficients from the regression of the return on human wealth on the marketable security returns. Defining $\mathbf{x}^e_i \equiv \mathbf{x}_i + \mathbf{b}_i$ as the vector of effective asset demands, we see from (17) that effective asset demands exhibit the standard separation property. This reflects the fact that, while the returns on human capital are not directly marketable, the component of the return which is linearly related to the returns on the marketable securities is indirectly marketable by appropriate offsetting positions in the marketable securities. The asset holdings of the individual may be represented as the sum of effective asset holdings $\mathbf{x}^e_i$ and an investment in the component of human wealth whose return is orthogonal to the returns on marketable assets. We refer to this as approximate portfolio separation since the first component exhibits portfolio separation, and the second component has no effect on the relative demands for marketable assets.

The Mayers model leads to an asset pricing equation which is identical to that of the Sharpe–Lintner model if the market portfolio is defined as the sum of the effective investment vectors $\mathbf{x}^e_i$.

Inflation and International Asset Pricing

Stochastic inflation has no effect on the foregoing results, provided that a common inflation rate can be defined for all investors and returns are restated in real terms. However, the international asset pricing models of Solnik (1974) and Stulz (1981) distinguish between nationalities precisely on the basis of their price indices, which may differ on account of either a violation of commodity price parity or differences in tastes and consumption baskets (see Adler and Dumas 1983).

Define $\widetilde{\pi}_i$ as the inflation rate in the numeraire currency for investor i. Then, to a high order of approximation, which becomes exact as the time interval approaches zero, the mean and variance of real wealth can be written as

$$\begin{aligned}\overline{W}_i = &\sum_j x_{ij}(\mu_j - r) \\ &+ W_{0i}\left(1 + r - \widetilde{\pi}_i + \sigma^i_{\pi\pi}\right) \\ &- \sum_j x_{ij}\sigma^j_{j\pi}\end{aligned} \tag{18}$$

$$\begin{aligned}S_i^2 = &\sum_j \sum_k x_{ij}x_{ik}\sigma_{jk} - 2W_{0i}\sum_j x_{ij}\sigma^i_{k\pi} \\ &+ W^i_{0i}\sigma^i_{\pi\pi}\end{aligned} \tag{19}$$

where W_{0i} is the investor's initial wealth.

The asset demand vector is then

$$x_i = \theta_i{}^{-1}\mathbf{\Omega}^{-1}(\boldsymbol{\mu} - r1) + \mathbf{b}_i \tag{20}$$

Where $\mathbf{b}_i \equiv W_{0i}\,\mathbf{\Omega}^{-1}\,\sigma^i_x$ is the vector of coefficients from the regression at the individual's aggregate inflation risk, $W_{0i}\widetilde{\pi}_i$, on security returns. If we compare (20) with (17), it is apparent that this international asset pricing model is isomorphic to the Mayers' non-marketable wealth model with individual inflation risks playing the same role as human capital.

Black (1974) has modelled segmentation in international capital markets by introducing a tax on foreign security holdings for residents of one country. This model is isomorphic to Brennan's (1970) tax model, if the foreign securities are thought of as paying dividends on which only domestic residents are taxable. Stulz (1981) extends Black's model by prohibiting negative taxes on short sales: as one might expect, this causes some indeterminacy in the pricing relations

since the marginal conditions of portfolio optimality are no longer always satisfied.

Intertemporal Models

Merton (1973) showed that the classical one-period CAPM can be extended to an intertemporal setting in which investors maximize the expected utility of lifetime consumption. With continuous trading and suitable restrictions on the stochastic process of asset prices, the essential mean-variance analysis is retained, the major innovation being that at each instant the individual may be represented as maximizing the expected utility of a derived utility function, defined over wealth and a set of S state variables describing the future investment and consumption opportunity sets. The state dependent derived utility function induces $(S + 1)$ fund separation in the risky asset portfolio, and the vector of risky asset demands may be written

$$x_1 = \theta_i^{-1}\mathbf{\Omega}^{-1}(\boldsymbol{\mu} - r1) - \sum_{s=1}^{S} \gamma_{is}\mathbf{\Omega}^{-1}\xi_s \quad (21)$$

where ξ_s is the vector of covariances of asset returns with the change in state variable S and γ_{is} depends on the utility function. Aggregation of asset demands and the imposition of the market clearing condition lead to an asset pricing equation in which asset risk premia are a linear function of covariances with aggregate wealth *and* covariances with changes in the state variables or factors that described the investment opportunity set. In the absence of prior information about the relevant state variables this model is empirically indistinguishable from the arbitrage pricing theory. Breeden (1979) showed that if consumption preferences are time separable this 'multi-beta' pricing model can be collapsed to a single beta measured with respect to changes in aggregate consumption, the 'consumption' CAPM(CCAPM), and much effort has been expended on testing this form of the model despite the difficulties of measuring consumption flows.

Campbell (1993) developed a model with recursive utility which, unlike the standard time-additive utility function defined over consumption, does not satisfy the von Neumann–Morgenstern axioms but does allow the intertemporal marginal rate of substitution to vary independently of risk aversion. This model contains elements of both the CAPM and the CCAPM in that expected returns depend on the covariances of asset returns with both consumption and the market return.

Recent Empirical Developments

During the 1990s renewed interest in Merton's (1973) 'intertemporal' CAPM (ICAPM) was generated by the empirical failures of both the CAPM and the CCAPM, the increasing evidence of time variation in investment opportunities, and the empirical success of an atheoretical three-factor model of security returns developed by Fama and French (FF) (1992, 1993) to account for high returns on small firms and the low returns on growth stocks relative to value stocks. The FF model could be interpreted as a version of either the APT or the ICAPM if no restrictions were placed on the types of factors that could enter these models. However, the factors that are important for pricing in the APT are those that explain the covariance of (one-period) returns, while the factors in the ICAPM are those that forecast future returns. Merton (1973) had suggested the interest rate as an example of an ICAPM state variable, and Nielsen and Vassalou (2006) showed formally that the only state variables that are relevant for the ICAPM are those with information about the current and future interest rate and the slope of the capital market line which is shown as rM in Fig. 1. Brennan et al. (2004) constructed a version of the ICAPM in which the interest rate and slope of the capital market line follow a joint Markov process, and showed that its empirical performance was at least as good as that of the FF model. Brennan and Xia (2006) used this framework to derive expressions for the prices of cash flow claims which depend explicitly on current capital market conditions as measured by the interest rate and the slope of the capital market line, as well as on the characteristics of the underlying cash flow. This implies that stock prices vary with discount rates as well as cash flow expectations, and Campbell

and Vuolteenaho (2004) showed that, if market betas are decomposed into components due to changes in cash flow expectations and to changes in discount rates, then risk premia are associated primarily with the cash flow component of beta. These models attribute the low returns on growth stocks to the greater proportion of their risk arising from discount rate changes.

The classic CAPM may hold even with time variation in investment opportunities. Constantinides (1980, 1982) has identified two sets of sufficient conditions for the simple CAPM to hold with a time varying interest rate. In his models the social investment opportunity set is stationary and consists only of risky investments: stochastic variation in the interest rate then does not affect the CAPM relation if there is either demand aggregation or full Pareto efficiency of asset markets. Either condition is sufficient for prices to be determined as though there existed a single representative individual; for such an individual stochastic variation in the interest rate is irrelevant since the interest rate represents only a shadow price and not a real investment opportunity. Finally, the single period nature of the CAPM is retained if individuals behave myopically, ignoring stochastic variation in the investment opportunity set: this occurs if and only if the utility function is logarithmic.

Time variation in the distribution of asset returns can affect tests of asset pricing models even if the CAPM is true. For example, if betas and risk premia are time varying, then average returns need not be related to average betas as predicted by the CAPM even if period by period returns and betas are. Lettau and Ludvigson (2001) argued that the predictive power of the CCAPM is considerably enhanced by allowing the covariances of asset returns to depend on a measure of the aggregate consumption–wealth ratio. However, Lewellen and Nagel (2006) argued that time variation in risk premia is unlikely to be sufficient to account for the observed value anomaly.

See Also

► Finance

Bibliography

Adler, M., and B. Dumas. 1983. International portfolio choice and corporation finance: A synthesis. *Journal of Finance* 38: 925–984.

Bergman, Y. 1985. Time preference and capital asset pricing models. *Journal of Financial Economics* 14: 145–159.

Black, F. 1972. Capital market equilibrium with restricted borrowing. *Journal of Business* 45: 444–455.

Black, F. 1974. International capital market equilibrium with investment barriers. *Journal of Financial Economics* 1: 337–352.

Breeden, D. 1979. An intertemporal asset pricing model with stochastic consumption and investment opportunities. *Journal of Financial Economics* 7: 265–296.

Brennan, M. 1970. Taxes, market valuation and corporate financial policy. *National Tax Journal* 23: 417–427.

Brennan, M., and Y. Xia. 2006. Risk and valuation under an intertemporal capital asset pricing model. *Journal of Business* 79: 1–35.

Brennan, M., A. Wang, and Y. Xia. 2004. A simple model of intertemporal capital asset pricing and its implications for the Fama–French three factor model. *Journal of Finance* 59: 1743–1776.

Campbell, J. 1993. Intertemporal asset pricing with consumption data. *American Economic Review* 83: 487–512.

Campbell, J., and T. Vuolteenaho. 2004. Good beta, bad beta. *American Economic Review* 94: 1249–1275.

Cass, D., and J. Stiglitz. 1970. The structure of investor preferences and asset returns, and separability in portfolio allocation: A contribution to the pure theory of mutual funds. *Journal of Economic Theory* 2: 122–160.

Chamberlain, G. 1983. A characterization of the distributions that imply mean- variance utility functions. *Journal of Economic Theory* 29: 185–201.

Cochrane, J. 2005. *Asset pricing theory.* Revised ed. Princeton: Princeton University Press.

Constantinides, G. 1980. Admissible uncertainty in the intertemporal asset pricing model. *Journal of Financial Economics* 8: 71–86.

Constantinides, G. 1982. Intertemporal asset pricing with heterogeneous consumers and without demand aggregation. *Journal of Business* 55: 253–267.

Cox, J., J. Ingersoll, and S. Ross. 1985. An intertemporal general equilibrium model of asset prices. *Econometrica* 5: 363–384.

Fama, E., and K. French. 1992. The cross-section of expected returns. *Journal of Finance* 47: 427–465.

Fama, E., and K. French. 1993. Common risk factors in the returns on stock and bonds. *Journal of Financial Economics* 33: 3–56.

Fisher, I. 1906. *The nature of capital and income.* New York: Macmillan.

Hicks, J.R. 1934a. A note on the elasticity of supply. *Review of Economic Studies* 2: 31–37.

Hicks, J.R. 1934b. Application of mathematical methods to the theory of risk. *Econometrica* 2: 194–195.

Lettau, M., and S. Ludvigson. 2001. Resurrecting the (C) CAPM: A cross-sectional test when risk premia are time-varying. *Journal of Political Economy* 109: 1238–1287.
Lewellen, J., and S. Nagel. 2006. The conditional CAPM does not explain asset pricing anomalies. *Journal of Financial Economics* 82: 289–314.
Lintner, J. 1965. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *Review of Economics and Statistics* 47: 13–37.
Markowitz, H. 1952. Portfolio selection. *Journal of Finance* 7: 77–91.
Markowitz, H. 1958. *Portfolio selection: Efficient diversification of investments*. New York: Wiley.
Mayers, D. 1972. Non-marketable assets and capital market equilibrium under uncertainty. In *Studies in the theory of capital markets*, ed. M. Jensen. New York: Praeger.
Merton, R. 1969. Lifetime portfolio selection under uncertainty: The continuous-time case. *Review of Economics and Statistics* 51 (3): 247–257.
Merton, R. 1972. An analytic derivation of the efficient portfolio frontier. *Journal of Financial and Quantitative Analysis* 7: 1851–1872.
Merton, R. 1973. An intertemporal capital asset pricing model. *Econometrica* 41: 867–887.
Nielsen, L.T., and M. Vassalou. 2006. The instantaneous capital market line. *Economic Theory* 28: 651–654.
Roll, R. 1977. A critique of the asset pricing theory's test; Part I: On past and potential testability of the theory. *Journal of Financial Economics* 4: 129–176.
Ross, S. 1976. The arbitrage theory of capital asset pricing. *Journal of Economic Theory* 13: 341–360.
Ross, S. 1978. Mutual fund separation in financial theory: The separating distributions. *Journal of Economic Theory* 17: 254–286.
Rubinstein, M. 1976. The valuation of uncertain income streams and the pricing of options. *Bell Journal of Economics* 7: 407–425.
Samuelson, P. 1970. The fundamental approximation theorem of portfolio analysis in terms of means, variances, and higher moments. *Review of Economic Studies* 37: 537–542.
Sharpe, W.F. 1964. Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance* 19: 425–442.
Solnik, B. 1974. An equilibrium model of international capital markets. *Journal of Economic Theory* 8: 500–524.
Stulz, R. 1981. A model of international asset pricing. *Journal of Financial Economics* 9: 383–406.
Tobin, J. 1958. Liquidity preference as behavior towards risk. *Review of Economic Studies* 25: 65–86.
Von Neumann, J., and O. Morgenstern. 1947. *Theory of games and economic behavior*. 2nd ed. Princeton: Princeton University Press.

Capital Budgeting

E. Solomon

A sub-field within economics and finance, the principal concern of capital budgeting is the optimal deployment of funds into capital expenditures. The mainstream of the field, developed essentially in the 1950s and 1960s, consists of two threads of inquiry. (1) How should a company measure the investment worth of a capital expenditure proposal? (2) How should a company set the minimum required rate of return for a capital expenditure proposal?

Historical Evolution

The concept of a *capital budget*, as opposed to an *operating budget*, originated in public finance. Many governments – the United States Federal government is a notable exception – have long maintained separate accounts for capital expenditures; that is, expenditures on capital assets that provide benefits over relatively long time periods.

In the private sector, separate budgeting for capital expenditure has an almost equally long history. The development, however, of the coherent body of thought now known as *capital budgeting* began only after World War II. Three forces drove that development:

(1) A dramatic postwar increase in private capital spending that led to increased interest in how such expenditures should be made.
(2) The postwar development of national income accounts which provided a vehicle for plausible economic projections, a necessary condition for rational choice.
(3) The publication in 1951 of two seminal books: *Capital Budgeting*, by Joel Dean, and *The Theory of Investment of the Firm*, by Friedrich and Vera Lutz.

Rational capital budgeting requires a correct basis for measuring the investment worth of each capital expenditure proposal.

Although capital expenditure decisions have long been regarded as one of the critical responsibilities of top-management (and, indeed, of Corporate Boards of Directors), before the 1950s the decisions themselves had been made on the basis either of intuitive judgements or poorly defined standards. Such inadequate approaches have been supplemented and, in some companies, supplanted by more robust and more quantitative criteria.

The Pay-Back Period

One of the earliest quantitative yardsticks used for assaying the investment worth of a capital expenditure proposal (and one that is still in use) is the project's pay-back period – the number of years required to recoup the initial outlay. Because the measure ignores the size and duration of benefits beyond the pay-back period itself, it is a poor proxy for 'profitability'. Nonetheless, it provides a quick screening device for rejecting some proposals as well as for selecting among alternative investment proposals that involve purchases of equipment having approximately equal lives.

The Average Rate of Profit

The earliest measure used for a project's expected profitability is the ratio of the average annual flow of profit expected from the project to the average investment dedicated to the project – both measured in conventional accounting terms. The measure has been increasingly discarded because it is a poor proxy for true profitability on two counts: (1) it ignores the *timing* of expected benefits, and (2) it is subject, both with respect to the numerator and the denominator, to the vagaries of depreciation accounting.

The DCF Rate of Return

The discounted cash-flow rate-of-return measure (hence DCF), which relates the incremental cash inflows attributable to a project to the incremental cash outlays required by it, has gradually supplanted the average-rate accounting measure. In principle, the DCF rate of return (or internal rate of return) is identical to the long-used financial measure for the effective yield to maturity on a bond; that is, it is the rate at which the present value of all incremental cash or equivalent benefits expected from an investment is equal to the incremental outlays required by that investment.

Net Present Value

If a company has a correct estimate for the rate of return that is required by the market on an investment with a given degree of riskiness, the DCF return offered by that investment proposal provides an infallible guide to whether or not it should be accepted. Exactly the same result can be achieved by an alternative process: if the present value of a project's net cash flows, discounted at its required rate of return, *exceeds* the present value of the outlays it entails, then the proposal should be accepted; that is, all proposals that have a positive *net* present value should be undertaken. Although both approaches yield the same correct result for accept–reject decisions, the net-present-value approach is a superior one in two special situations. (1) Some investment proposals (especially those designed to accelerate cash-inflows) have more than one DCF rate of return solution. In such cases (i.e. those with two or more positive solutions), none is a correct measure of the project's expected profitability. (2) When more than one proposal is acceptable by either standard, but only one can be executed because the two are mutually exclusive, the net-present-value approach invariably provides a better answer to the 'which is better?' question.

The Required Rate of Return

Both the DCF approach and the Net-Present-Value approach to investment decisions require a correct estimate of the required rate of return (or applicable discount rate) on the investment outlay that is being assayed. A number of increasingly sophisticated approaches to the estimation of that rate (also known as the appropriate 'cost of capital') have been developed. All such measures are now based on observable rates of return demanded in the marketplace by holders of the debt and equity securities that jointly finance the assets of the corporation. This rate, adjusted up or down for any differential riskiness of the particularly project that is being assayed, is now widely used as the 'hurdle' that any proposal must pass in order to be acceptable.

The rationale is a straightforward one. An investment that yields a *higher* rate of return than the market-determined cost of the funds it requires, has a positive net present value; that is, it creates wealth for the owners as well as for society as a whole. That is how Adam Smith's 'invisible hand' gets translated into practice.

See Also

- ► Investment decision criteria
- ► Present value

Bibliography

Bierman, H., and S. Smidt. 1960. *The capital budgeting decision*. New York: Macmillan.

Dean, J. 1951. *Capital budgeting*. New York: Columbia University Press.

Lutz, F., and V. Lutz. 1951. *Theory of investment of the firm*. Princeton: Princeton University Press.

Robichek, A. (ed). 1967. *Financial research and management decisions*. New York: Wiley.

Solomon, E. 1959. *The management of corporate capital*. Glencoe: Free Press . Contains an extensive bibliography of the literature up to 1959.

———. 1962. *The theory of financial management*. New York: Columbia University Press.

Capital Controls

Kristin J. Forbes

Abstract

Capital controls can take many different forms and are broadly defined as any restrictions on the movement of capital across a country's borders. This article focuses on the debate on the merits of capital controls for emerging markets and developing economies. It describes the potential costs and benefits of capital controls, focusing on the recent empirical literature evaluating the impact of capital controls.

Keywords

Bretton Woods system; Bubbles; Capital account liberalization; Capital controls; Distortions; Dutch disease; *encaje* (Chile); Fixed exchange rates; Foreign direct investment; International monetary fund; Keynes, J. M.; Kindleberger, C.; Monetary policy; Nurkse, R.; Portfolio investment; Risk diversification

JEL Classifications

F21

Capital controls are any restrictions on the movement of capital into or out of a country. Capital controls can take a wide variety of forms. For example, capital controls can be quantity-based or price-based, or apply to only capital inflows, only capital outflows, or all types of capital flows. Capital controls can also be directed at different types of capital flows (such as at bank loans, foreign direct investment or portfolio investment) or at different types of actors (such as at companies, banks, governments or individuals).

Most developed countries believe that the benefits from the free movement of capital across borders outweigh the costs, and therefore have very limited (if any) capital controls in place today. For emerging markets and developing economies, however, there has been a long-

standing debate on the desirability of capital controls. Assessing the impact of capital controls is complicated due to a number of factors, including the various forms in which they can be structured. This article discusses the recent debate on capital controls, focusing on the theoretical arguments for and against controls and the existing empirical evidence on their impact.

History of the Debate

Throughout the 20th century, economists have regularly expressed concerns about international capital flows. For example, in the 1940s Ragnar Nurkse worried about 'destabilizing capital flows' and in the 1970s Charles Kindleberger described the role of capital in driving 'manias, panics and crashes' (see Nurkse 1944; Kindleberger 1978). When the world's leading economies met at Bretton Woods in 1944 to formulate rules governing the international financial system, John Maynard Keynes and other delegates debated the role of capital controls. The resulting compromise required that members of the International Monetary Fund (IMF), one of the newly created international monetary institutions, allow capital to be freely exchanged and convertible across countries for the purpose of all current account transactions, but permitted members to implement capital controls for financial account transactions. Most countries had capital controls in place at this time.

Over the following years, however, many developed countries gradually removed their capital controls, so that by the 1980s most had few controls in place. In the early and mid-1990s, many emerging markets and developing countries also began to lift their capital controls. The impact initially appeared to be positive – capital flowed into countries with liberalized capital accounts, investment and growth increased, and asset prices rose. In fact, support for lifting capital controls was so widespread that in 1996–7 leading policymakers discussed amending the rules agreed to at Bretton Woods to extend the IMF's jurisdiction to include capital movements and make capital account liberalization a goal of the IMF. In mid-1997, however, a series of financial crises started in Asia and spread across the world, appearing to disproportionately affect emerging markets that had recently liberalized their capital accounts. This series of crises sparked a reassessment of the desirability of capital controls for emerging markets and developing economies.

In a sharp sea change, many leading policymakers and economists began to support the use of capital controls for emerging markets in some circumstances, especially taxes on capital inflows. Much of this support was based on the belief that controls on capital inflows could reduce a country's vulnerability to financial crises. From 2002 to 2005, several emerging markets (such as Colombia, Russia and Venezuela) also implemented new controls on capital inflows, largely to reduce the appreciations of their currencies. Over the same period, however, several large emerging markets (such as India and China) moved in the opposite direction and lifted many of their existing controls.

Benefits and Costs of Capital Controls

The free movement of capital across borders can have widespread benefits. Capital inflows can provide financing for high-return investment, thereby raising growth rates. Capital inflows – especially in the form of direct investment – often bring improved technology, management techniques, and access to international networks, all of which further raise productivity and growth. Capital outflows can allow domestic citizens and companies to earn higher returns and better diversify risk, thereby reducing volatility in consumption and income. Capital inflows and outflows can increase market discipline, thereby leading to a more efficient allocation of resources and higher productivity growth. Implementing capital controls can reduce a country's ability to realise these multifaceted benefits.

On the other hand, the free movement of capital across borders can also have costs. Countries reliant on foreign financing will be more vulnerable to 'sudden stops' in capital inflows, which can cause financial crises and/or major currency depreciations. Large volumes of capital inflows

can cause currencies to appreciate and undermine export competitiveness, causing what is often called the 'Dutch disease'. The free movement of capital can also complicate a country's ability to pursue an independent monetary policy, especially when combined with a fixed exchange rate. Finally, capital inflows may be invested inefficiently due to a number of market distortions, thereby leading to overinvestment and bubbles that create additional challenges. Capital controls could potentially reduce these costs from the free movement of capital.

Empirical Evidence on Capital Controls

Since capital controls can have costs and benefits, evaluating the desirability and aggregate impact of capital controls is largely an empirical question. (See Eichengreen 2003, on the potential costs and benefits of capital controls.) Not surprisingly, an extensive literature has attempted to measure and assess the effects of capital controls.

The most studied experience with capital controls is the Chilean *encaje* – a market-based tax on capital inflows from 1991 to 1998 so structured that the magnitude of the tax decreased with the maturity of the capital flow. Chile's experience with capital controls is generally viewed positively, largely due to Chile's strong economic performance during the period the controls were in place. Empirical studies of the impact of Chile's capital controls, however, have reached several general conclusions. First, there is no evidence that the capital controls moderated the appreciation of Chile's currency (which was the primary purpose of the capital controls). Second, there is little evidence that the controls protected Chile from external shocks. Third, there is some evidence that the controls raised domestic interest rates (at least in the short term). Fourth, there is some evidence that the controls did not affect the volume of capital inflows, but did lengthen the maturity of capital inflows. Finally, the capital controls significantly raised the cost of financing for small and medium-sized firms and distorted the mechanisms by which Chilean companies procured financing. The general conclusion from this work is that Chile's strong economic performance during the 1990s resulted from sound macroeconomic and financial policies, not the capital controls, and that the capital controls had both costs and benefits. (See Forbes 2007, for more information on this literature and the Chilean capital controls.)

A second major branch of literature examining the impact of capital controls focuses on the effects of lifting capital controls (that is, capital account liberalization). The majority of this work uses macroeconomic data, typically focusing on how capital account liberalization raises economic growth using cross-country growth regressions. Prasad et al. (2003) is a detailed survey of this literature and shows that, although several papers find a robust, positive effect of capital account liberalization on growth, other papers find no significant effect, and most papers find mixed evidence. This literature is generally read as showing weak evidence that lifting capital controls may have some positive effect on growth.

There are several explanations for the inconclusive results in this macroeconomic literature assessing the impact of capital controls. First, it is extremely difficult to measure capital account openness and to capture the various types of capital controls in a simple measure that can be used for empirical analysis. Second, different types of capital flows and controls may have different effects on growth and other macroeconomic variables. For example, controls on portfolio investment may be more beneficial than other types of capital controls. Third, the impact of removing capital controls could depend on a range of other factors that are difficult to capture in cross-country regressions, such as a country's institutions, financial system, corporate governance or even the sequence in which different controls are removed. Fourth, capital controls can be very difficult to enforce (especially for countries with undeveloped financial markets) so the same capital control may have different degrees of effectiveness in different countries. Finally, most countries that remove their capital controls undertake simultaneously a range of reforms and undergo structural changes, so that it can be difficult to isolate the impact of removing the controls.

(For additional details on the challenges in measuring the impact of capital controls, see Eichengreen 2003; Forbes 2006; Magud and Reinhart 2006; Prasad et al. 2003.)

Given these challenges in measuring the impact of capital controls, it is not surprising that the empirical literature has had difficulty documenting their effects on growth at the macroeconomic level. To put these results in perspective, however, the current status of this literature is similar to the literature in the 1980s and 1990s on how trade liberalization affects economic growth. Economists generally believe that trade openness raises growth, but most of the initial work on this topic also focused on cross-country, macroeconomic studies and reached inconclusive results. At a much earlier date, however, several papers using microeconomic data and case studies found compelling evidence that trade liberalization raises productivity and growth.

Similarly, recent work based on microeconomic data has been much more successful than the macroeconomic literature in documenting the effects of capital controls. Forbes (2006) surveys this new literature, which covers a variety of countries and periods, uses a range of approaches and methodologies, and builds on several different fields. This literature has, to date, reached five general results. First, capital controls reduce the supply of capital, raise the cost of financing, and increase financial constraints – especially for smaller firms and firms without access to international capital markets. Second, capital controls reduce market discipline in financial markets and the government, leading to a more inefficient allocation of capital and resources. Third, capital controls distort decision-making by firms and individuals as they attempt to minimize the costs of the controls, or even evade them outright. Fourth, the effects of capital controls vary across different types of firms and countries, reflecting different pre-existing economic distortions. Finally, capital controls can be difficult and costly to enforce, even in countries with sound institutions and low levels of corruption. Therefore, this series of microeconomic studies suggests that capital controls have widespread and pervasive costs, but has not yet provided significant evidence of the benefits of capital controls.

Conclusions

The debate on the effects and desirability of capital controls is likely to continue and to motivate new academic research. Most economists agree that countries should gradually lift their capital controls as they grow and develop, and that developed countries should have few (if any) capital controls in place. Most economists also believe that the free movement of capital can have widespread benefits, but that in countries with weak financial systems, poorly developed institutions, and vulnerable macroeconomies the free movement of capital can also generate distortions and increase a country's vulnerability. As a result, emerging markets and developing countries that currently have capital controls should work to address the shortcomings in their economies as they liberalize their capital accounts. There continues to be widespread disagreement, however, on the exact sequencing of these reforms and the optimal pace of capital account liberalization for emerging markets and developing economies.

See Also

- ▶ International Capital Flows
- ▶ International Monetary Institutions
- ▶ Kindleberger, Charles P. (1910–2003)
- ▶ Nurkse, Ragnar (1907–1959)

Bibliography

Eichengreen, B. 2003. *Capital flows and crises*. Cambridge, MA: MIT Press.

Forbes, K. 2006. The microeconomic evidence on capital controls: No free lunch. In *Capital controls and capital flows in emerging economies: Policies, practices and consequences*, ed. S. Edwards. Chicago: University of Chicago Press.

Forbes, K. 2007. One cost of the Chilean capital controls: Increased financial constraints for smaller traded firms. *Journal of International Economics.*

Kindleberger, C. 1978. *Manias, panics and crashes: A history of financial crises*. New York: Wiley.
Magud, N., and C. Reinhart. 2006. *Capital controls: An evaluation*. Working Paper No. 11973. Cambridge, MA: NBER.
Nurkse, R. 1944. *International currency experience: Lessons of the interwar experience*. Geneva: League of Nations.
Prasad, E., K. Rogoff, S.-J. Wei, and M. Kose. 2003. *Effects of financial globalization on developing countries: Some empirical evidence*. Occasional Paper No. 220. Washington, DC: IMF.

Capital Flight

Brendan Brown

This term describes the phenomenon of funds fleeing across the national frontier in search of greater safety. The driving forces behind capital flight include actual or feared monetary instability, confiscatory taxation, war and revolution. Examples of the phenomenon can be found through several centuries. A low level of liquidity and high costs of international communication at first limited the potential scope of capital flight. The earliest 'modern' example was the largescale movement of French funds to London during the Franco-Prussian war. Capital flight has reached in the twentieth century a frequency and importance previously unseen.

The first major episode was the flight of capital during World War 1 out of France, Italy and the Central Powers, into the neutral countries – principally Switzerland, the Netherlands, and Sweden. The capital movements were 'accommodated' to a large extent by speculators in the neutrals buying the belligerent currencies at big discounts to their theoretical gold pars in the hope that large gains would be made once peace was restored.

Defeat brought a new outpouring of capital from Central Europe. Funds fled the Austro-Hungarian crown out of fear that the Successor States would 'nationalize' crowns on their own territory – blocking a substantial share of private holdings and insisting on tax-registration before converting the remainder into the new national money. At first, buyers of the Austro-Hungarian notes could be found in Italy's new territories (acquired from Austria–Hungary) in the expectation (correct) that the Italian authorities would ultimately convert its new subjects' holdings into liras at a favourable rate. Then buyers appeared in the form of tourists attracted in swarms to Vienna during 1920–1921 by the cheap crown.

The flight out of the mark was at first driven by fear that hurge taxes would be levied by the new Republic to meet the internal and external costs of defeat. After a brief respite in the last three-quarters of 1920, capital flight got new impetus from the gathering reparations crisis. The German government was suspected of deliberately inflating to demonstrate the 'impossibility' of paying reparations, whilst the danger of a French invasion of the Ruhr increased. Germany was again the source of huge capital flight in the years 1929–1931, driven this time by the spectre of political instability (from mid-1929, the Nazi vote in elections rose strongly) and of national bankruptcy.

The next major episode of capital flight was from France. The fascist riots in February 1934, then the prospects of a 'Front Populaire' government coming to power (May 1935) and of a large devaluation of the franc to reflate the economy, unleashed a huge outflow of funds. The formation of a Centre–Centre Right government under Daladier in spring 1938 marked the turning point. In the next 18 months, funds returned to France despite the growing menace of war. For Britain and the European neutrals (Holland, Belgium and Switzerland), by contrast, spring 1938 marked the start of a period of capital flight as funds sought refuge in the USA. There were three great waves and a final smaller wave between mid-1938 and the end of 1939: autumn 1938 (Munich crisis), spring 1939 (German occupation of Prague), August 1939 (Nazi-Soviet Pact and invasion of Poland) and November 1939 (feared invasion of the Low Countries). The Bank of England financed the outflows by undertaking massive dollar sales.

Capital flight changed direction dramatically as soon as France sued for an armistice (June

1940). Investors in Axis Europe and in the remaining European neutrals (particularly Switzerland) feared that the USA would freeze their funds and these were transferred into Swiss francs or to Latin America. For the next decade, Switzerland was the principal recipient of refuge funds – which in the early post-war years came largely from France. The USA was not regarded as a safe-haven – not just because of its wartime freeze of most foreign assets, but also because of the cooperation of the US authorities with European governments in securing the repatriation of flight capital.

In general, the postwar industrial world has not been struck by the huge waves of capital flight driven by political fears which marked the years 1914–1940. The mid and late 1970s were a period of large movements of flight capital, but these were driven primarily by inflation. The inflows to Switzerland from France, Italy and Britain in 1976 reflected largely the high inflation in these countries and the non-indexation of their tax structures (particularly with respect to capital). During 1978, the spectre of high and rising inflation in the USA caused international funds to flee the dollar. Just when inflation fears began to moderate following the turn in US monetary policy of October 1979, a new fillip was given to capital outflows from the USA by the freezing of Iranian assets. Investors in much of the Third World, particularly OPEC, feared that if revolution brought to power a government unfriendly to Washington, their dollar assets might not be safe.

In almost all the episodes of capital flight mentioned, foreign investors and creditors have played a disproportionately large role. Foreign capital is less tied down by 'convenience factors'. Domestic residents in general have less to lose than foreigners from the introduction of exchange restrictions. Whereas foreigners might not be able to buy anything with frozen balances (except, perhaps, tourist services), residents would be able to use their funds freely on a normal range of goods – albeit possibly curtailed by import controls.

A general property of capital flight driven by fears of future disaster is that it occurs in waves, not continuously. The wave-like motion reflects discontinuous changes in the probability of the possible 'bad state of the world' becoming reality. News – a frequent cause of shifts in probability assessments – is by its nature sudden. Alarming new information causes investors to revise upwards the share of hedge-assets (usually foreign) in their portfolio. During the period of portfolio-adjustment a wave of capital flight becomes apparent. Once adjustment is complete, the wave subsides. Under a floating exchange rate system, the waves are sublimated into abrupt fluctuations in currency values. The exchange rate falls to a point where investors see sufficient return on holding the 'troubled' money at the margin to delay re-arranging their portfolio. As trade flows respond to the exchange rate change, the portfolio adjustment begins to take place.

Capital flight can reach such a force as to cause national bankruptcy (meaning that foreign credits are frozen and exchange restrictions introduced). For example, the official foreign exchange reserves may have become exhausted; foreign loans be impossible to obtain; interest rate rises (which in principle might stem capital outflows) be infeasible because they would intensify deflation, increasing the risk of domestic political tumult or bank failures; a downward float of the currency be ineffective in strengthening the capital account because it gives rise to a wage–price spiral or invites retaliation by other nations concerned with 'unfair' competition in trade. Governments sometimes pre-empt a forced bankruptcy by coming to a 'voluntary' re-scheduling arrangement with foreign creditors and imposing a range of controls on domestic capital exports. Such measures are costly. The country's credit-rating would be adversely affected for decades to come. A tradition of economic and political liberalism might well be damaged irreparably. In some respects, a liberal government which prevents its citizens from protecting their wealth against the coming to power of a dictatorship or against a foreign invader is already in league with the enemy.

The fear of forced bankruptcy is not the only motive for government to seek to limit capital

flight. Measures may be introduced as a 'sop' to labour when economic policy is being tightened. Alternatively, the authorities may hope that the measures will raise the level of domestic investment and employment of real wages. A reduced degree of capital flight should mean that interest rates can settle at a lower level. In general, though, measures against capital flight are largely ineffective, unless policed by methods inconsistent with a liberal society. Traffic in banknotes is one obvious loophole, especially where the given country has land frontiers and is a tourist centre. Other loopholes include false invoicing in trade and compensation payments.

Such transactions often lie behind the large negative 'errors and omissions' items in balance of payments statistics for countries susceptible to capital flight. They may also be responsible for the positive 'errors and omissions' for the countries receiving flight capital. The positive errors could reflect foreign hoarding demand for the domestic currency (for example, Swiss franc notes accumulated outside Switzerland) or inflows of flight capital being hidden behind domestic names for fear of freezing (for example, much of the inflows to the USA in 1939–1940 were disguised behind US names and gave rise to a large positive errors item in the US balance of payments at that time).

Measures against capital flight might indeed increase its extent. Domestic investors would realize that they could not quickly raise the proportion of foreign currency in their portfolio. Hence, if the political and economic climate at home worsened, they could be 'underprotected' for a long time. To hedge this possibility, they might painstakingly via available loopholes accumulate foreign holdings to a level higher than justified simply by present risks.

See Also

- ▶ Exchange Control
- ▶ Hot Money
- ▶ International Capital Flows

Bibliography

Blankart, C. 1919. *Die Devisenpolitik während des Weltkrieges*. Zurich: Orell Fussli.
Gutmann, I. 1913. *Das Französische Geldwesen im Kriege 1870–78*. Strassborg: Trubner.
Koeppel, W. 1931. *Kapitalflucht*. Berlin: Wilhelm Christians.

Capital Gains and Losses

E. Malinvaud

Abstract

How capital gains and losses are distinct from income raises subtle and unresolved issues. Whereas national accountants measure income as the sum of the value of production and net current transfers, thus excluding stock revaluations that change the level of wealth, Hicks's definition implies that expected stock revaluations count as income. Such revaluations due to inflation benefit net debtors but mean losses for households. Irreversible environmental damage and depletion of non-renewable resources are often treated as capital loss, but great uncertainty affects the estimation of consequences, rendering the emergence of an objective methodology for economic decisions is particularly difficult.

Keywords

Capital gains and losses; Capital gains taxation; Comprehensive definition of income; Depreciation; Exhaustible resources; Expected and unexpected capital gains or losses; Fisher, I.; Haig–Simons definition of income; Hicks, J. R.; Income, definition of; Inflation; Intergenerational equity; National accounting; Residential real estate; Uncertainty

JEL Classifications

E22

National accounting has made the definition of capital gains and losses rather precise in practice,

but fundamentally their distinction from income raises quite subtle issues, about which great economists have long been wavering. Whenever it becomes important, inflation gives to some of these issues a fresh relevance. Much remains to be learned, moreover, on how capital gains affect economic behaviour and how the allocation of resources ought to deal with the capital losses resulting from current activity.

Definition

Although the reference books such as United Nations (1969) are not explicit enough about this basic notion, national accounting systematically applies the following

$$\Delta W = Y + CT + CG - C \qquad (1)$$

where ΔW is the variation of wealth between the beginning and end of the period under consideration, Y is income, CT the net capital transfer received (gifts, bequests, capital taxes and subsidies), CG the net capital gain and C consumption. The identity applies to any agent or group of agents. This identity may be taken as the de facto definition of net capital gains (that is, gains *minus* losses), to the extent that well- defined rules are used for the flows Y, C and CT, which appear in the current accounts, and to the extent that wealth is assumed to be unambiguously determined.

Looking carefully at the existing rules, one, however, realizes that the distinction between income and net capital gain is conventional to a large extent. It is precisely on the choice of this convention that some important questions about the definition of incomes lie.

Chapter 7 of Fisher (1906) shows that defining the concept of income was not an easy task for economists. Fisher's own preferred definition, 'the services of capital', may not seem quite clear, but it can be identified with consumption. This would make the whole of investment belong to capital gains, a solution that was seriously discussed by Samuelson (1961) but has hardly any advocate today. At the other extreme, the 'comprehensive definition of income', also called the Haig–Simons definition, was proposed by economists studying income taxes (Haig 1921; Simons 1938); income would be equal to the sum of consumption and wealth increase, thus leaving neither capital gains, nor capital transfers in Eq. 1. One now most commonly refers to the definition introduced by Hicks (1939 p. 172), 'A man's income is the maximum value which he can consume during a week, and still expect to be as well off at the end of the week as he was at the beginning'.

National accountants, however, measure income as the sum of the value of production and net current transfers. Production is essentially computed from physical outputs and inputs, valued at current prices and aggregated. This means that stock revaluations that explain part of the change of wealth are not incomes but capital gains or losses. Hicks's definition, on the contrary, implies that expected stock revaluations belong to income. In Eq. 1 only windfalls would be true capital gains. But whether the change of value of an asset should be classified as expected or not is most often not clear. (How long in advance should it have been expected? Should an outside observer be able to make sure that the asset holder had expected the change?) The distinction between expected and unexpected capital gains or losses, however, remains essential in economic analysis.

Inflation

The most sizeable asset revaluations result from changes of the price level. When inflation is important, a good proportion of these revaluations are, moreover, expected by all agents. Their occurrence then plays a role in the determination of the equilibrium of all exchanges and economic operations, inducing in particular high interest rates. On the other hand, the change of nominal wealth becomes of little interest in comparison with the change of real wealth; 'real capital gains' should then be distinguished from nominal ones. Hence, inflation perturbs the significance of normal accounting rules; new measurements are required for correct assessments of income flows (Jump 1980).

This applies first to business accounting, in which reference to historical costs underestimates physical assets and depreciation of fixed capital, while it overestimates net returns from financial assets. This explains the search for new or alternative accounting rules that would be better suited in cases of fast inflation and would more correctly draw the line between income and capital gains or losses. This search went as far as the stage of implementation in the United Kingdom (see Walton 1978).

At the level of the whole economy, when the rules of national accounting are applied, real capital gains and losses resulting from variations of the general level of prices are important. Typically they benefit enterprises and government, which are net debtors, whereas they mean large losses for households. When all these capital gains and losses are imputed to incomes, on the ground that they must have been expected, the current accounts of firms and government appear substantially more favourable, whereas sizeable redistribution is also found as between groups of households (see Bach and Stephenson 1974; Babeau 1978; Wolff 1979).

The question has been considered whether national account practices should not be revised so as to better record true incomes in times of inflation (see Hibbert 1982). A prerequisite is the regular production of national balance sheets. When this is done, important capital gains and losses, due for instance to booms in real estate or share prices, also appear beyond those due to changes of the general price level.

Capital Gains in Economic Behaviour

Most econometric studies tend to neglect capital gains as flows, although wealth and indebtedness are often taken into account. The role of capital gains on the consumption behaviour of households has, however, been studied. Up to now the results have been rather inconclusive (Bhatia 1972; Peek 1983; Pesaran and Evans 1984).

In all likelihood the difficulty comes from the fact that some capital gains are purely transitory, whereas most of them have some degree of permanence, but this degree varies widely from one to the other. A pure windfall is comparable to an exceptional gift; accidental losses or war damages occur once for all, whereas capital losses due to an inflation that is expected to last may appear to be as permanent as interest incomes, even sometimes as wage incomes. But to classify capital gains according to their supposed permanence is far from being an obvious operation.

Gains on the value of corporate shares have a permanent component following from the firms' policy of retaining part of their profits. This is why increases of retained earnings have been considered as likely to increase household consumption, but not as much as an increase of permanent income would, since the size of undistributed profits varies a good deal with business conditions (Feldstein and Fane 1973; Malinvaud 1986).

The problem becomes still more complex when capital gains are correlated with cost changes for items of household wealth. An extreme case occurs when prices of residential real estate increase: owners of houses make a capital gain, but simultaneously the cost of housing increases by the corresponding amount; whether houses are let or used by their owners, a stimulating effect on real consumption is doubtful.

Capital Losses, Conservation and Welfare

The existence of capital gains and losses raises a number of issues for the theory of allocation of resources, for instance what should be the taxation of capital gains (David 1968; Green and Sheshinski 1978), or how best to organize insurance against capital losses. But particular attention nowadays concerns the damages that economic activity causes to the environment and to reserves of exhaustible resources (Fisher 1981).

Not all environmental effects mean capital losses; many of them are just externalities in the normal course of economic activity. But irreversible damages to the forests, the soil or even the climate must also be recognized and are usually not recorded as consumption or as inputs to

production. Depletion of non-renewable reserves is similarly often treated as capital loss.

The detrimental effects of many of these losses will appear mainly in a rather distant future. Whether or not losses should be accepted – what for instance should be the optimal speed of depletion of natural resources – raises difficult questions of intergenerational equity, on which economists have uncomfortably to enter the field of social philosophy.

The problem cannot be discarded here on the ground that proper discounting makes the distant future negligible. Indeed, in the purest case, the shadow discounted price of an exhaustible resource is as high in the future as it is now, for as long as the resource will remain used (Hotelling 1931). The remote future must then be taken into account for present decisions.

It is moreover notorious that enormous uncertainties affect the purely physical estimation of the consequences involved. Neither the effects of carbon dioxide emission on the climate, nor the existing reserves of fossil fuels, nor the future emergence of appropriate technologies for the wider use of renewable energy can be securely assessed. Under such circumstances, the emergence of an objective methodology for economic decisions is particularly difficult.

Bibliography

Babeau, A. 1978. The application of the constant price method for evaluating the transfer related to inflation: The case of French households. *Review of Income and Wealth* 24: 391–414.

Bach, G., and J. Stephenson. 1974. Inflation and the redistribution of wealth. *The Review of Economics and Statistics* 56: 1–13.

Bhatia, K. 1972. Capital gains and the aggregate consumption function. *American Economic Review* 62: 866–879.

David, M. 1968. *Alternative approaches to capital gains taxation*. Washington, DC: Brookings Institution.

Feldstein, M., and G. Fane. 1973. Taxes, corporate dividend policy and personal savings: The British experience. *The Review of Economics and Statistics* 55: 399–411.

Fisher, I. 1906. *The nature of capital and income*. New York: Macmillan.

Fisher, A. 1981. *Resource and environmental economics*. Cambridge: Cambridge University Press.

Green, J., and E. Sheshinski. 1978. Optimal capital-gains taxation under limited information. *Journal of Political Economy* 86: 1143–1158.

Haig, R. 1921. The concept of income: Economic and legal aspects. In *The federal income tax*, ed. R. Haig. New York: Columbia University Press.

Hibbert, J. 1982. *Measuring the effects of inflation on income, saving and wealth*. Paris: OECD.

Hicks, J. 1939. *Value and capital*. 2nd ed. Oxford: Oxford University Press. 1946.

Hotelling, H. 1931. The economics of exhaustible resources. *Journal of Political Economy* 39: 137–175.

Jump, G. 1980. Interest rates, inflation expectations, and spurious elements in measured real income and saving. *American Economic Review* 70: 990–1004.

Malinvaud, E. 1986. Pure profits as forced saving. *Scandinavian Journal of Economics* 88: 109–130.

Peek, J. 1983. Capital gains and personal saving behaviour. *Journal of Money, Credit and Banking* 15: 1–23.

Pesaran, M., and R. Evans. 1984. Inflation, capital gains and UK personal savings: 1953–81. *Economic Journal* 94: 237–257.

Samuelson, P. 1961. The evaluation of 'social income': Capital formation and wealth. In *The theory of capital*, ed. F. Lutz and D. Hague. London: Macmillan.

Simons, H. 1938. *Personal income taxation*. Chicago: University of Chicago Press.

United Nations. 1969. *A system of national accounts*. New York: United Nations.

Walton, J. 1978. Current cost accounting: Implications for the definition and measurement of corporate income. *Review of Income and Wealth* 24: 357–390.

Wolff, E. 1979. The distributional effects of the 1969–75 inflation on holdings of household wealth in the United States. *Review of Income and Wealth* 25: 195–208.

Capital Gains Taxation

William Gentry

Abstract

Capital gains taxation is the taxation of gains or losses from owning assets, usually as part of an income tax. Typically, tax systems measure capital gains or losses upon realization so that capital gains are taxed only when assets are sold. These realization-based tax rules create a number of behavioural incentives. Investors have an incentive to maximize the value of tax

deferral by delaying the sale of assets. Capital gains taxes can also affect incentives for investing in risky assets. The realization-based tax rules also complicate the estimation of the revenue consequences of changing the tax rate on capital gains.

Keywords

Capital gains and losses; Capital gains taxation; Inflation; Tax base; Tax incentives for saving; Tax planning; Taxation of corporate profits

JEL Classifications

H2

Capital gains taxation involves the taxation of changes in asset values, usually in the context of an income tax rather than as a separate tax. Under a pure income tax, these gains or losses would be measured on a periodic basis (for example, annually) and would be adjusted for inflation. However, actual tax systems tend to deviate in several important ways from this hypothetical treatment. The most important of these deviations is that capital gains are typically measured upon the realization of the gain or loss rather than under accrual accounting. The taxation of capital gains creates a wide variety of incentive issues, especially given the deviations between their tax treatment under a pure income tax and their treatment under actual tax rules.

Administrative Issues

While the concept of a capital gain or loss from the ownership of an asset is straightforward, administering a tax on capital gains is a complicated part in the income tax codes of most countries. The primary difficulty arises from the challenge of measuring the size of a capital gain or loss over a specified period of time. This difficulty has led to most capital gains being taxed upon realization rather than as they accrue. The exceptions to this general rule tend to be for relatively sophisticated investors (for example, brokers) on assets that are relatively liquid and easily valued (for example, publicly traded equities). Realization-based taxation means that taxpayers keep records of the purchase price of assets, known as the basis in the asset, and calculate the gain or loss as the difference between the sales price and this basis when the asset is sold. The basis in an asset can be adjusted over time, with the most common type of adjustment being for the depreciation allowances accorded to depreciable assets.

An important issue in measuring capital gains is whether the gain is adjusted for changes in purchasing power created by inflation. Countries vary in their treatment of capital gains created by inflation. Most countries include the portion of the gain that is due to inflation in the tax base, but a few countries allow the asset's tax basis to be adjusted for inflation so that the tax base includes only the real portion of the capital gain. A pure income tax would allow for an adjustment for inflation, but such an adjustment would be part of a system that adjusted all forms of capital taxation for inflation.

In many countries, capital gains face lower marginal tax rates than other sources of income. Two rationales motivate these lower tax rates. First, policymakers may want to encourage investment in activities that generate capital gains. Second, the preferential tax rates provide an ad hoc method of adjusting tax burdens for inflation in tax systems that do not index the measurement of capital gains for inflation. These preferential rates, which can include the exemption of capital gains from income taxation, often depend on meeting a minimum holding period (for example, preferential rates apply to 'long-term' capital gains that are earned on assets held for longer than one year) and may apply only to specific types of assets (for example, gains on corporate stock qualify for preferential tax rates but gains on collectibles do not).

Another cumbersome feature of capital gains taxation is the specific rules dealing with how gains and losses offset each other. Typically, these loss-offset provisions limit a taxpayer's ability to use capital losses to offset other sources of income. The motivation for these limitations is that realization-based taxation provides taxpayers with the option of deferring the tax on gains but

accelerating the deductions for losses through a strategy of holding on to appreciated assets but selling assets with losses.

In terms of administration, Auerbach (1991) and Auerbach and Bradford (2004) propose tax systems that allow for realization-based tax rules that would mimic the incentive and revenue effects of accrual taxation of capital gains. Such tax reforms would eliminate many of the complicated incentive effects created by current administrative rules for capital gains taxation.

Incentive Effects

Taxing capital gains creates a variety of incentive, or disincentive, effects. Since taxing capital gains is typically part of a broader regime to tax capital income, the tax on capital gains can affect incentives to save. As a tax on capital income, the capital gains tax reduces the return to saving, which can have a theoretically ambiguous effect on the level of savings in the economy. Of course, since many countries provide preferential tax treatment for capital gains compared with other forms of capital income, tax policy towards capital gains often increases the return to saving by reducing the effective tax rate on savings compared with a regime without preferential tax rates for capital gains.

Capital gains taxation can also affect incentives for taking risk. A tax on capital gains from risky investments reduces the expected return to these investments, which one might expect would discourage investment in risky assets. However, the tax on capital gains also reduces the variance in the payoffs to investing in risky assets and this reduction in variance may encourage investors to increase their investments in risky assets. The net effect of the reduction in both the expected return and the variance in returns may actually imply that the theoretical effect of a higher tax rate on capital gains is an increase in the amount of risk taking (see Domar and Musgrave 1944). This result, however, rests on the symmetric tax treatment of gains and losses. When loss offset rules are imperfect, such that gains face a higher marginal tax rate than losses, then the theoretical predictions are much more complicated and it becomes more likely that the capital gains tax reduces the amount of risk taking in the economy because gains face a higher tax rate than losses.

The relative tax treatment of capital gains and other forms of capital income can also affect investors' portfolio choices (see Poterba 2002; Poterba and Samwick 2002). If capital gains face lower effective tax rates, due to either preferential tax rates or the ability to defer taxes by deferring realization of income, investors may prefer to invest in assets that are likely to generate capital gains rather than assets that generate interest or dividend income. In addition to affecting portfolio decisions, the relative tax treatment of different forms of capital income may also affect relative asset prices and expected returns (see Klein 1999).

The realization-based feature of capital gains taxation creates several tax planning incentives (see Stiglitz 1983). By not selling an appreciated asset, an investor can postpone paying the tax liability on the associated capital gain. This deferral of taxation reduces the discounted value of the tax (assuming that the statutory tax rate will remain constant in the future). This incentive to delay the realization of capital gains is known as the 'lock-in' effect since the tax liability that would be triggered by selling an asset reduces the incentive for investors to sell appreciated assets and locks them into holding assets. In the United States, the incentive to defer the realization of capital gains is compounded by tax rules that allow heirs to step-up the basis of appreciated assets that they inherit, which eliminates the income tax on capital gains on bequeathed assets.

In addition to incentives to delay the realization of capital gains, realization- based taxation also creates an incentive to accelerate the realization of capital losses since these losses can reduce taxation on other types of income (though this offset is possibly limited by loss offset rules) or capital gains on other assets (see Constantinides 1983; Poterba 1987; Auerbach et al. 2000). This pattern of selective realization leads to the tax planning advice that taxpayers should sell their losers and hold their winners. In essence, realization-based taxation provides taxpayers with an option of whether to pay taxes, and it is typically more

advantageous to exercise this option for assets that have lost value.

While most of the incentives discussed above deal with decisions made by investors, the tax treatment of capital gains can also affect the supply of different assets. For example, corporations may alter their payout policies in response to the relative tax treatment of dividends and capital gains. To the extent that capital gains face a lower effective tax rate than dividends at the investor level, corporations have an incentive to retain earnings so that investors can recognize income as capital gains rather than distribute earnings as dividends. Retaining earnings due to this tax rate differential does not necessarily imply that it leads to an increase in corporate investment. Instead of increasing investment, corporations that eschew dividends can repurchase shares as an alternative mechanism to distribute cash to shareholders (see Green and Hollifield 2003). These share repurchases allow investors to time their tax liabilities since the decision to sell shares back to the firm is discretionary and, for the shareholders who sell, the income associated with the transaction faces capital gains tax rates rather than dividend tax rates.

Revenue Consequences

One of the more contentious issues surrounding capital gains taxation is the effects of capital gains taxes on government revenues. From the government's perspective, the incentive effects discussed above create opportunities for lost revenue. While the overall revenue effect of capital gains taxation depends on the whole myriad of incentives discussed above, much of the empirical literature on this issue has focused on the capital gains realization decisions of individuals. An important empirical issue has been separating how capital gains realizations respond to short-run fluctuations in the tax rate (or anticipated changes in tax rates) from how long-term realizations behaviour responds to the tax rate (or the 'permanent' response to tax changes). Auerbach (1988) examines the time series evidence in the United States and documents a large timing response of capital gains to anticipate tax rate changes but finds limited evidence of a permanent response of capital gains realizations to tax rates. Burman and Randolph (1994) examine a panel of US household taxpayers; their results also point towards a much larger transitory response than permanent response to changes in capital gains tax rates. Taken together, these studies cast doubt on the claim that reductions in capital gains tax rates can be self-financing.

See Also

- ▶ Capital Gains and Losses
- ▶ Individual Retirement Accounts
- ▶ Taxation of Corporate Profits
- ▶ Taxation of Income

Bibliography

Auerbach, A.J. 1988. Capital gains taxation in the United States: Realizations, revenue, and rhetoric. *Brookings Papers on Economic Activity* 2: 595–631.

Auerbach, A.J. 1991. Retrospective capital gains taxation. *American Economic Review* 81: 167–178.

Auerbach, A.J., and D.F. Bradford. 2004. Generalized cash-flow taxation. *Journal of Public Economics* 88: 957–980.

Auerbach, A.J., L.E. Burman, and J.M. Siegel. 2000. Capital gains taxation and tax avoidance: New evidence from panel data. In *Does Atlas shrug? The economic consequences of taxing the rich*, ed. J. Slemrod. New York: Russell Sage Foundation.

Burman, L.E., and W.C. Randolph. 1994. Measuring permanent responses to capital-gains tax changes in panel data. *American Economic Review* 84: 794–809.

Constantinides, G.M. 1983. Capital market equilibrium with personal tax. *Econometrica* 51: 611–636.

Domar, E.D., and R.A. Musgrave. 1944. Proportional income taxation and risk-taking. *Quarterly Journal of Economics* 58: 387–422.

Green, R.C., and B. Hollifield. 2003. The personal-tax advantage of equity. *Journal of Financial Economics* 67: 175–216.

Klein, P. 1999. The capital gain lock-in effect and equilibrium returns. *Journal of Public Economics* 71: 355–378.

Poterba, J.M. 1987. How burdensome are capital gains taxes? Evidence from the United States. *Journal of Public Economics* 33: 157–172.

Poterba, J.M. 2002. Taxation, risk-taking, and household portfolio behavior. In *Handbook of public economics*, vol. 3, ed. A.J. Auerbach and M.S. Feldstein. Amsterdam: North-Holland.

Poterba, J.M., and A.A. Samwick. 2002. Taxation and household portfolio composition: U.S. evidence from the 1980s and 1990s. *Journal of Public Economics* 87: 5–38.

Stiglitz, J.E. 1983. Some aspects of the taxation of capital gains. *Journal of Public Economics* 21: 257–294.

Capital Goods

Harald Hagemann

Capital goods are a series of heterogeneous commodities, each having specific technical characteristics. Outside the hypothetical case where real capital consists of a single commodity, it is impossible to express the stock of capital goods as a homogeneous physical entity. As a consequence of capital's heterogeneous nature its measurement has become the source of many controversies in the history of economic thought.

The function of capital goods is production. Unlike labour ('in the raw') and (non-cultivated) land, capital goods are not given, they are themselves produced. Being an output as well as an input, the size and variation of the capital stock are intra-economic phenomena. Because real capital is not an 'original' factor of production but is the result of economic processes in which it participates as one of the determinants, the formation of real capital or investment is the central channel through which all other determinants, be they technical progress, changes in labour supply or the exploitation of natural resources, influence the long-run development of an industrial system.

A distinction is normally made between durable or fixed capital, including not only plant and machinery but also buildings and other essential parts of the industrial infrastructure which are used up only partially during the year, and circulating capital, consisting of stocks of raw materials, semi-finished goods, etc., capital which is fully used up during the production period and must therefore be replaced in full.

Capital has at least two different aspects: capital as goods and capital as value. From a technological point of view, produced means of production are a condition for the operation of any social and economic system, once Smith's early and rude state of society is overcome. It was Marx who emphasized that these necessary physical instruments of production become 'capital' only under the capitalistic rules of the game when the means of production are separated from the labourers and owned by the capitalists. Thus the means of production possess a double aspect in capitalistic societies: on the one hand 'capital' is understood to mean the total of heterogeneous goods and equipment designed for specific uses (productive concept), on the other hand it is regarded as a homogeneous fund of value and source of 'unearned' income in the form of profits (portfolio concept).

The value of the capital goods corresponding to each system of production, even with a constant technique, will change with income distribution whichever the unit in which they are measured. Current relative prices change when the rate of profits or the real wage rate changes, so that the same physical capital represents a different value whereas different stocks of capital goods can have the same value. Furthermore, only in long-run equilibrium will a given stock of capital goods have the same value whether it be determined as the accumulated sum of past investment expenditures or as the expected future net returns discounted back to the present at the ruling rate of profits.

Another way of measuring capital goods is in terms of labour time directly and indirectly required to produce them, the appropriately dated quantities of labour compounded at the various given rates of profits. As the analyses of Joan Robinson (1956), who called it 'real capital', and Sraffa (1960) show, it is impossible to get any notion of capital as a measurable quantity independent of distribution and prices.

Whereas the individual is concerned with the extent to which he owns capital goods as a store of wealth and a source of income, society as a whole is never faced with problems of buying or selling capital goods against money or credit. Greater output unambiguously requires a greater amount of capital goods, given the degree of capacity

utilization and technology. These additional capital goods can be provided only by a process of accumulation or net investment.

Emphasis on the strategic role of the capacity to produce capital goods in the domestic economy plays a decisive role in the analyses of Fel'dman (1928) and Lowe (1955, 1976). Both authors take as their starting point Marx's famous two-departmental scheme of expanded reproduction, modifying it in an adequate way to include all activities that increase the capacity of an economy to produce output in one sector. During the Soviet industrialization debate in the late twenties, Fel'dman formalized the notion that investment-priority for the capital-goods sector was a precondition for attaining a higher growth rate. Structural incapacity to supply enough capital goods will prevent a rise in the saving ratio from being fully transformed into the desired level of investment. But it has to be taken into account that a one-sided preoccupation with this 'Fel'dman constraint' on the investment capacity side may bring the 'Preobrazhenski constraint' on the consumption side into action. If the initial capacity of the capital goods industry is just sufficient to replace the worn-out machines, growth can only take place as a result of a temporary reduction in the output of consumer goods which may be impossible for subsistence reasons. In this case a *circulus vitiosus* will emerge.

The strategic role of the machine tools sector and the compulsion to enlarge first the equipment in capital goods industries were also dealt with by economists discussing the growth and planning problems of underdeveloped countries in the Fifties and Sixties (see, for example, Dobb 1960 and Mathur 1965). Countries like India which lack a self-sufficient machine tools sector can speed up their transformation process by foreign trade. The Fel'dman constraint would be binding only if the domestic output of machine tools could not be supplemented with imports.

The perception that there is a group of fixed capital goods which hold the strategic position in any industrial system like seed corn for agricultural production, led Lowe to the conclusion that it is useful to split up the capital goods sector in the Marxian scheme of reproduction into two subsectors. In his 'tripartite' scheme of three vertically integrated sectors, the first produces primary equipment goods or 'machine tools' which are directly used for production in sectors I and II. Sector II produces the secondary equipment goods which are used as inputs only in sector III producing consumer goods, which means that the capital stock in the latter is not transferable. Thus sector I is the only one capable not only of producing machines for other sectors but also for itself; it is therefore a *self-reproducible sector.* In Sraffa's terminology, sector I represents the 'basic system'.

The sub-division of the capital goods group is relevant for investigating the structural conditions for steady growth and, even more, in addressing questions of 'traverse analysis', when the problem of structural change is moved to the centre of the stage. The decisive problem that the economy faces upon departing from a steady growth path is the inadequacy of the old capital stock. The dynamic traverse from one steady growth path to another necessarily involves a change in the whole quantity structure, especially the rebuilding of the capital stock. The economy cannot change output unless it first changes inputs, i.e. the capital goods group must provide the commodities demanded for changing the inputs to produce the new output pattern. The production of machine tools is the bottle-neck which any process of rapid expansion must overcome. The key to a higher growth rate lies in increasing the shares of sector I. The same logic requiring that the system as a whole first has to change inputs before it can change output makes such an increase dependent on the prior expansion of the capital stock of this sector. Whereas in the two-sectoral Fel'dman model this is only possible by a policy of putting a larger proportion of new machine tools into the production of more machine tools, in the Lowe model an additional *ex post* transfer of machine from sector II to sector I is possible, thereby shortening the time of adjustment. Both models come to the same result, namely that in order to increase the growth rates of total output and consumption output in the long run, at first a temporary fall in the growth rate of consumption output is necessary.

The neo-Austrian theory developed by Hicks is characterized by a completely different treatment of the durable means of production. In his neo-Austrian model, a stream of labour inputs is converted into a stream of final outputs (consumption goods). 'Capital goods are simply stages in the process of production' (Hicks 1973, p. 5), i.e. they are regarded as intermediate products which don't appear explicitly but are implied and produced within each process of 'maturing' of original inputs into the final product. Thus the intertemporal aspect of production and consumption is placed into the forefront of the analysis; time is the essence of capital in the Austrian view. By treating fixed capital as if it were working capital, Hicks does not recognize the need for a special machine-tools sector. There is no basic product in this model. Hence, the production process is not 'circular'; the neo-Austrian approach turns out to be a further variant of the production theoretic paradigm of marginalist analysis, which conceives of the production process as a 'one-way avenue that leads from "Factors of production" to "Consumption goods"' (Sraffa 1960, p. 93).

It is precisely the focus on the adjustment problems caused by the impact of technical innovations that has led Hicks to his vertical representation of the productive structure. In contrast to Leontief–Sraffa–Lowe systems, in Hicks neither intersectoral transactions, nor therefore the effects of innovation upon industrial structure, are shown. Hicks sees the decisive advantage of the Austrian method in its ability to cope with the important fact that process innovations nearly always involve the introduction of new capital goods. This would lead to insurmountable difficulties in the traverse analysis if capital goods were physically specified because 'there is no way of establishing a physical relation between the capital goods that are required in the one technique and those that are required in the other' (Hicks 1977, p. 193). A similar explanation is given by Pasinetti who develops his theory of structural change in terms of vertically integrated sectors. While conceding that the input–output model gives more information on the structure of an economic system at any point in time, he points out that because of the change of input–output coefficients and the 'breaking down' of the inter-industry system over time, the vertically integrated model is superior for dynamic analysis (see Pasinetti 1981, pp. 109–17). Measuring capital goods in units of vertically integrated productive capacity of the final commodity 'has an unambiguous meaning through time, no matter which type of technical change, and how much of it, may occur' (p. 178).

Whilst it is true that a sectorally disaggregated approach encounters difficulties when the effects of innovations connected with the introduction of new capital goods are studied, the price that Austrian-type models have to pay for their linear 'imperialism' is rather high. Technical change takes place at the industry level, a characteristic which is completely washed out in vertically integrated models. The industry-specific nature of technical change also implies that, contrary to Pasinetti's assumption, rates of productivity growth in the different vertically integrated sectors cannot be thought of as being independent of each other. How could the new capital goods be produced without the old ones existing at the beginning of the traverse? Thus the existence of a basic system remains relevant, even when the basic product(s) is(are) changing its(their) quality. Innovations introducing new consumption goods cannot be dealt with in a satisfactory way. All this does not imply that the concept of vertically integrated sectors is meaningless, on the contrary, it can be very helpful as a complementary perspective. But it illustrates that input–output models emphasizing intersectoral interdependencies retain conceptual priority.

Fixed capital has two other important dimensions: its degree of capacity utilization, and its durability. Thus the choice of cost-minimizing technique involves the choice of the 'planned' degree of capital utilization and the choice of the economic lifetime of a fixed capital good. The latter can best be dealt with on the basis of a von Neumann–Sraffa treatment of fixed capital goods (which contains Hicks's neo-Austrian model as a special case) as a joint part of the gross output, thus identifying machines of different ages as different commodities. To every technically possible lifetime corresponds a specific w-r relation

which may slope upwards over some range for a given truncation (in which case the prices of partly worn-out machines become negative and premature truncation is advantageous), whereas the w–r frontier is always downwards sloping. The analysis of the choice of the optimal lifetime or truncation period shows that with constant or increasing efficiency the maximum technical lifetime will always be chosen, independently of income distribution. With decreasing or changing efficiency, however, premature truncation may become profitable (see Hagemann and Kurz 1976). A change in the wage rate (rate of profits) will generally lead to changes in the optimal economic lifetimes of fixed capital goods. With more complex patterns of the time profile of efficiency, the return of the same truncation period at different intervals of the rate of profits is possible, a phenomenon closely linked to reswitching of techniques (see also Schefold 1974).

See Also

- ▶ Accumulation of Capital
- ▶ Capital as a Factor of Production

References

Dobb, M. 1960. *An essay on economic growth and planning*. London: Routledge.

Fel'dman, G.A. 1928. On the theory of growth rates of national income, I and II. In *Foundations of soviet strategy for economic growth*, ed. N. Spulber. Bloomington: Indiana University Press, 1964.

Hagemann, H., and H.D. Kurz. 1976. The return of the same truncation period and reswitching of techniques in neo-Austrian and more general models. *Kyklos* 29(4): 678–708.

Harcourt, G.C. 1972. *Some Cambridge controversies in the theory of capital*. Cambridge: Cambridge University Press.

Hicks, J. 1973. *Capital and time. A neo-Austrian theory.* Oxford: Clarendon Press.

Hicks, J. 1977. *Economic perspectives. Further essays on money and growth*. Oxford: Clarendon Press.

Lowe, A. 1955. Structural analysis of real capital formation. In *Capital formation and economic growth*, ed. M. Abramovitz. Princeton: Princeton University Press.

Lowe, A. 1976. *The path of economic growth*. Cambridge: Cambridge University Press.

Mathur, G. 1965. *Planning for steady growth*. Oxford: Blackwell.

Pasinetti, L.L. 1981. *Structural change and economic growth. A theoretical essay on the dynamics of the wealth of nations*. Cambridge: Cambridge University Press.

Robinson, J. 1956. *The accumulation of capital*. London: Macmillan.

Schefold, B. 1974. Fixed capital as a joint product and the analysis of accumulation with different forms of technical progress. In *Essays on the theory of joint production*, ed. L.L. Pasinetti. London: Macmillan.

Sraffa, P. 1960. *Production of commodities by means of commodities. Prelude to a critique of economic theory.* Cambridge: Cambridge University Press.

Capital Measurement

W. Erwin Diewert and Paul Schreyer

Abstract

Capital measures provide an indicator of wealth and of capital services, the contribution of assets to production. The wealth stock is the market value of assets, whereas capital services are measured in proportion to the quantity of past investment, adjusted for the relative efficiency of different vintages and capital goods in production. Although the two measures of capital are different, they are derived from a single theoretical framework whose centrepiece is a fundamental equilibrium relationship between stocks and flows of capital. Index number theory is used to guide the empirical implementation of stock and flow measures.

Keywords

Aggregation; Asset inflation; Asset pricing; Asset values; Böhm-Bawerk, E.; Capital measurement; Capital services; Capital utilization; Depreciation; Fixed base (chain) indexes; Geometric (declining balance) depreciation model; Index numbers; Interest, market rates; Linear efficiency decline model; One-hoss shay efficiency model; Price and quantity indexes; Producer equilibrium; Production function; Rates

of return; Rental price; Stocks and flows; Straight line depreciation model; User cost; Value ratios; Walras, L.

JEL Classifications
C43; E01; E22; O47

Capital measures are constructed for two main purposes: (1) to measure wealth (the market value of assets) and (2) to analyse the role of capital in production. Because capital is durable, the value of using it in any given year is not the same as the value of owning it. There are thus different measures of capital depending on the purpose of accounting. However, these different measures should be consistently derived from a single framework.

The scope of the discussion below is restricted to fixed assets and land; we do not deal with financial or intangible assets, inventories or environmental assets.

Fundamental Relations Between Stocks and Flows of Capital

In equilibrium, the *stock value* of an asset is equal to the discounted stream of future rental payments for *capital services* that the asset is expected to yield, an insight that goes at least back to Walras (1874) and Böhm-Bawerk (1888).

Let the price of an n-period old asset purchased at the beginning of period t be P_n^t. When prices change over time, it is necessary to distinguish between the observable rental prices for the asset at different ages in period t and future expected rental prices. Let f_n^t be the rental price of an n-period old asset at the beginning ofzperiod t. Then the fundamental equation relating the stock value of an asset, P_n^t, to the sequence of rental prices by age, $\{f_n^t : n = 0,\ 1,\ 2, \ldots\}$ is:

$$\begin{aligned} P_n^t = {} & f_n^t + [(1+i^t)/(1+r^t)]f_{n+1}^t \\ & + [(1+i^t)/(1+r^t)]^2 f_{n+2}^t \\ & + [(1+i^t)/(1+r^t)]^3 f_{n+3}^t + \ldots \\ n = 0,\ & 1,\ 2, \ldots \end{aligned} \tag{1}$$

where the i_n^t are expected rates of change of rental prices that are formed at the beginning of period t. For simplicity, it has been assumed that i_n^t does not depend on the asset's age. The term $1 + r^t$ is the discount factor that makes a dollar received at the beginning of period t equivalent to a dollar received at the beginning of period $t + 1$. Thus, the r_n^t are one-period *nominal interest rates* where the assumption has been made that the term structure of interest rates is constant. However, as the period t changes, r^t and i^t can change. The sequence of stock prices $\{P_n^t\}$ is not affected by general inflation provided that it affects the expected asset inflation rates i_n^t and the nominal interest rates r_n^t in a proportional manner.

The rental prices $\{f_n^t\}$ are potentially observable. In producer equilibrium, the ratio of any pair of rental prices equals the relative marginal productivity of the corresponding capital goods; see Hulten (1990).

By successive insertion for different P_n^t, (1) can be transformed into:

$$P_n^t = f_n^t + [(1+i^t)/(1+r^t)]P_{n+1}^t \tag{2}$$

or

$$\begin{aligned} f_n^t &= (1+r^t)^{-1}\left[P_n^t(1+r^t) - (1+i^t)P_{n+1}^t\right] \\ &= \left[P_n^t - (1+i^t)P_{n+1}^t/(1+r^t)\right];\ \ n \\ &= 0,\ 1,\ 2, \ldots \end{aligned} \tag{3}$$

Christensen and Jorgenson (1969) derived a version of (3) for the geometric depreciation model and end-of-period rental payments. Other variants are due to Christensen and Jorgenson (1973), Diewert (1980, 2005), Jorgenson (1989), Hulten (1990) and Diewert and Lawrence (2000).

(3) represents the *rental price* or *user cost* of an n-year old asset: the cost of using it during a period is given by the difference between the purchase price at the beginning of the period P_n^t and the value of the depreciated asset $(1 + i^t)P_{n+1}^t = P_{n+1}^{t+1}$ at the end of period t. Since this offset to the initial expense will be received only by the end of the period, it must be divided by the discount factor $(1 + r^t)$.

Depreciation, Asset Prices and User Costs

Depreciation is typically defined as the decline in asset value as one goes from an asset of a particular age to the next oldest at the same point in time; see Hicks (1939), Hulten and Wykoff (1981a, b), Hulten (1990), Jorgenson (1996) and Triplett (1996). Define the *depreciation rates* δ_n^t for an asset that is n periods old at the start of period t as:

$$\delta_n^t \equiv 1 - \left[P_{n+1}^t / P_n^t\right]; \quad n = 0, \ 1, \ 2, \ldots \tag{4}$$

Thus, given $\left\{P_n^t\right\}$, the period t sequence of $\left\{\delta_n^t\right\}$ is determined. Conversely, given $\left\{\delta_n^t\right\}$ and the price of a new asset in period t, $\left\{P_n^t\right\}$ is determined.

$$P_n^t = \left(1-\delta_0^t\right)\left(1-\delta_1^t\right)\ldots\left(1-\delta_{n-1}^t\right)P_0^t; \quad n = 0, \ 1, \ 2, \ldots \tag{5}$$

With expressions (5) and (3), the sequence of user costs $\left\{f_n^t\right\}$ can be expressed in terms of the price of a new asset at the beginning of period t, P_0^t, and $\left\{\delta_n^t\right\}$:

$$\begin{aligned} f_n^t &= (1+r^t)^{-1}\left(1-\delta_0^t\right)\ldots\left(1-\delta_{n-1}^t\right) \\ &\left[(1+r^t)-(1+r^t)\left(1-\delta_n^t\right)\right]P_0^t \\ &= (1+r^t)^{-1}\left[r^t+\delta_n^t(1+i^t)-i^t\right]P_n^t \quad n = 0, \ 1, \ 2, \ldots \end{aligned} \tag{6}$$

Thus, given any one of these sequences, all of the other sequences are completely determined. This means that assumptions about depreciation rates, the pattern of user costs by age or the pattern of asset prices by age cannot be made independently of each other. This point was first explicitly made by Jorgenson and Griliches (1967, 1972).

Aggregation

Asset prices are relevant for the construction of *wealth measures* of capital, and the user costs are relevant for the construction of *capital services measures*. Let there be N different types of assets and let the quantity of period t investment in asset i be I_i^t with a sequence of asset prices $\left\{P_{n,i}^t\right\}$. Then the value of the period t wealth stock is:

$$\begin{aligned} W_i^t \equiv{}& P_{0,i}^t \ I_i^{t-1} + P_{1,i}^t \ I_i^{t-2} + P_{2,i}^t \ I_i^{t-3} \\ &+ \ldots \ i \\ ={}& 1, \ 2, \ldots, N. \end{aligned} \tag{7}$$

To turn to capital services (we set aside issues of capital utilization), the flow of services that an asset of a particular age delivers is proportional to the corresponding quantity of past investment. The value of capital services for all ages of a given asset class i during period t using the sequence of user costs $\left\{f_{n,i}^t\right\}$ is:

$$\begin{aligned} S_i^t \equiv{}& f_{0,i}^t \ I_i^{t-1} + f_{1,i}^t \ I_i^{t-2} + f_{2,i}^t \ I_i^{t-3} \\ &+ \ldots \ i \\ ={}& 1, \ 2, \ldots, N. \end{aligned} \tag{8}$$

The value aggregates W_i^t and S_i^t can be decomposed into separate price and quantity components by standard index number methods, if each new unit of capital lasts only a finite number of periods, L. Define the period t price, user cost and quantity vectors, P_i^t, f_i^t and K_i^t respectively, as follows:

$$\begin{aligned} P_i^t &\equiv \left[P_{0,i}^t, \ P_{1,i}^t, \ldots, P_{L-1,i}^t\right]; f_i^t \\ &\equiv \left[f_{0,i}^t, \ f_{1,i}^t, \ \ldots, f_{L-1,i}^t\right]; K_i^t \\ &\equiv \left[I_i^{t-1}, \ I_i^{t-2}, \ldots, I_i^{t-L-1}\right]; \ i \\ &= 1, 2, \ldots, N. \end{aligned} \tag{9}$$

Fixed base or chain indexes may be used to decompose value ratios into price-change and quantity-change components. The values of W_i^t and S_i^t relative to their values in the preceding period, W_i^{t-1}, S_i^{t-1} have the following index number decomposition:

$$\begin{aligned} W_i^t / W_i^{t-1} = {}& P_i^W\left(P_i^{t-1}, \ P_i^t, \ K_i^{t-1}, \ K_i^t\right) \\ & Q_i^W\left(P_i^{t-1}, \ P_i^t, \ K_i^{t-1}, \ K_i^t\right); \\ i = {}& 1, \ 2, \ldots, N. \end{aligned} \tag{10}$$

$$S_i^t/S_i^{t-1} = P_i^S\left(f_i^{t-1},\ f_i^t,\ K_i^{t-1},\ K_i^t\right) \quad Q_i^S\left(f_i^{t-1},\ f_i^t,\ K_i^{t-1},\ K_i^t\right); \quad i = 1, 2, \ldots, N. \tag{11}$$

where P_i^W, P_i^S and Q_i^W, Q_i^S are bilateral price and quantity indexes respectively. In particular, Q_i^S measures the service flow of type i assets into production. It is thus an appropriate measure of capital input.

A functional form has to be chosen. For empirical work, Diewert (1976, 1992) has shown that the *Fisher* (1922) *ideal price and quantity indexes* appear to be 'best' from the axiomatic viewpoint, and can also be given strong economic justifications. The above index number approach to aggregating over vintages of capital was first suggested by Diewert and Lawrence (2000) and it is more general than the usual aggregation procedures for homogenous assets, which essentially assume that the different ages of the same capital good are perfectly substitutable so that linear aggregation techniques can be used.

However, most researchers use an index number approach to form price and quantity aggregates across *different* types of assets. The overall values of the period t wealth stock and capital services are respectively:

$$W^t \equiv P_1^{W,t}Q_1^{W,t} + P_2^{W,t}Q_2^{W,t} + P_3^{W,t}Q_3^{W,t} + \ldots \tag{12}$$

$$S^t \equiv P_1^{S,t}Q_1^{S,t} + P_2^{S,t}Q_2^{S,t} + P_3^{S,t}Q_3^{S,t} + \ldots \tag{13}$$

Akin to (10)–(11), the value aggregates W^t and S^t can be decomposed into separate price and quantity components. Define the period t price and quantity vectors, $P^{W,t}$, $P^{S,t}$ and $K^{W,t}$, $K^{S,t}$ respectively, as follows:

$$P^{W,t} \equiv \left[P_1^{W,t},\ P_2^{W,t}, \ldots, P_N^{W,t}\right]; P^{S,t} \equiv \left[P_1^{S,t},\ P_2^{S,t}, \ldots, P_N^{S,t}\right];$$
$$K^{W,t} \equiv \left[K_1^{W,t},\ K_2^{W,t}, \ldots, K_N^{W,t}\right]; K^{S,t} \equiv \left[K_1^{S,t},\ K_2^{S,t}, \ldots, K_N^{S,t}\right] \tag{14}$$

The values of W^t and S^t relative to their values in the preceding period, W^{t-1} and S^{t-1}, have the following index number decomposition:

$$W^t/W^{t-1} = P^W\left(P^{W,t-1},\ P^{W,t},\ K^{W,t-1},\ K^{W,t}\right) \quad Q^W\left(P^{W,t-1},\ P^{W,t},\ K^{W,t-1},\ K^{W,t}\right); \tag{15}$$

$$S^t/S^{t-1} = P^S\left(P^{S,t-1},\ P^{S,t},\ K^{S,t-1},\ K^{S,t}\right) \quad Q^S\left(P^{S,t-1},\ P^{S,t},\ K^{S,t-1},\ K^{S,t}\right); \tag{16}$$

where P^W, P^S and Q^W, Q^S are bilateral price and quantity indexes respectively. In particular, Q^S measures the overall service flow of capital into production.

Empirical Determination of Rates of Return and Asset Price Changes

Rates of return r^t can be based either on a balancing procedure or on market interest rates. The balancing procedure postulates that the value of capital services is equal to the value of gross operating surplus as shown by the national accounts plus the capital income of the self-employed. A rate of return is then chosen so that this equality holds. If market interest rates are used, there is still a choice between *ex ante* and *ex post* rates. Most empirical work on capital services has relied on an *ex post* balancing procedure based on Jorgenson and Griliches (1967, 1972) and Christensen and Jorgenson (1969). However, empirical problems arise when these methods yield highly volatile and sometimes negative user costs of capital. The debate has therefore continued – see Harper et al. (1989), Diewert (1980, 2005) and Schreyer (2006).

Possibilities for the choice of the asset inflation rates i^t include using the *ex post* asset price changes (consistent with the *ex post*, balancing procedure for rates of return), forecasting *ex ante* rates on the basis of *ex post* rates and assuming that expected asset price changes are equal to general inflation. The latter implies that the term $r^t - i^t$ in the user cost expression (6) becomes a real rate of return that is simple to measure and typically not too volatile. At the same time, the procedure may induce a bias in

C

user costs and capital measures if the prices of different assets move with different trends and/or if asset prices move very differently from general inflation.

Empirical Determination of Rates of Depreciation

Possibilities for determining depreciation rates include a number of approaches. First, information on market prices of assets of different age at the same point in time can be used to derive measures of depreciation. Empirical studies include Hall (1971), Beidelman (1973), Hulten and Wykoff (1981a, b) and Oliner (1996). The literature has been reviewed by Hulten and Wykoff (1996) and Jorgenson (1996). The second approach uses rental prices for assets where they exist, along with information on the rate of return and on asset prices to solve the user cost Eq. (6) for the rate of depreciation; for a review see Jorgenson (1996). The third approach is based on production function estimation where output is regressed on non-durable inputs and past investment. The estimated coefficients of the investment variable can be used to identify a constant rate of depreciation. Empirical studies using this approach include Epstein and Denny (1980), Pakes and Griliches (1984), Nadiri and Prucha (1996) and Doms (1996). The fourth method relies on insurance and other expert appraisals.

The fifth method makes assumptions about the relative efficiency sequence $\{f^t_n/f^t_0\}$ and the service life of assets, and then derives, via (1) and (5), a consistent measure of the rate of depreciation. For example, the *one-hoss shay* model of efficiency states that an asset yields a constant level of services throughout its useful life of L years: $f^t_n/f^t_0 = 1$ for $n = 0, 1, 2, \ldots, L-1$ and zero for $n = L, L+1, L+2, \ldots$. Another example is a *model of linear efficiency decline*, where the sequence $\{f^t_n/f^t_0\}$ is given by $f^t_n/f^t_0 \equiv [L-n]/L$ for $n = 0, 1, 2, \ldots, L-1$ and zero for $n = L, L+1, L+2, \ldots$.

The sixth method makes direct assumptions about the depreciation sequence $\{P^t_n/P^t_0\}$. The most frequent approaches are the *straight line depreciation model* and the *geometric* or *declining balance model*. Under the former, there is a constant amount of depreciation between every vintage: $P^t_n/P^t_0 = [L-n]/L$ for $n = 0, 1, 2, \ldots, L$ and zero for $n > L$. Under the latter, which dates back to Matheson (1910), there is a constant rate of depreciation $\delta^t_n = \delta$ for $n = 0, 1, 2, \ldots$. The geometric model greatly simplifies the algebra of capital measurement and has been supported empirically through studies on used asset markets; see Hulten and Wykoff (1981a, b). When there is only information on the average asset life L, the double declining balance method determines the rate of depreciation as $\delta = 2/[L+1]$.

See Also

- Capital Asset Pricing Model
- Capital Theory
- Depreciation
- Total Factor Productivity

Bibliography

Beidelman, C. 1973. *Valuation of used capital assets*. Sarasota: American Accounting Association.

Böhm-Bawerk, E. 1888. *The positive theory of capital*, trans. W. Smart. New York: G.E. Stechert, 1891.

Christensen, L., and D. Jorgenson. 1969. The measurement of U.S. real capital input, 1929–1967. *Review of Income and Wealth* 15: 293–320.

Christensen, L., and D. Jorgenson. 1973. Measuring the performance of the private sector of the U.S. economy, 1929–1969. In *Measuring economic and social performance*, ed. M. Moss. New York: Columbia University Press.

Diewert, W. 1976. Exact and superlative index numbers. *Journal of Econometrics* 4: 115–145.

Diewert, W. 1980. Aggregation problems in the measurement of capital. In *The measurement of capital*, ed. D. Usher. Chicago: University of Chicago Press.

Diewert, W. 1992. Fisher ideal output, input and productivity indexes revisited. *Journal of Productivity Analysis* 3: 211–248.

Diewert, W. 2005. Issues in the measurement of capital services, depreciation, asset price changes and interest rates. In *Measuring capital in the new economy*, ed. C. Corrado, J. Haltiwanger, and D. Sichel. Chicago: University of Chicago Press.

Diewert, W., and D. Lawrence. 2000. Progress in measuring the price and quantity of capital. In *Econometrics*

and the cost of capital: Essays in honor of Dale W. Jorgenson, ed. L. Lau. Cambridge, MA: MIT Press.
Doms, M. 1996. Estimating capital efficiency schedules within production functions. *Economic Inquiry* 34: 78–92.
Epstein, L., and M. Denny. 1980. Endogenous capital utilization in a short run production model: Theory and empirical application. *Journal of Econometrics* 12: 189–207.
Fisher, I. 1922. *The making of index numbers*. London: Macmillan.
Hall, R. 1971. The measurement of quality change from vintage price data. In *In Price indexes and quality change*, ed. Z. Griliches. Cambridge. MA: Harvard University Press.
Harper, M., E. Berndt, and D. Wood. 1989. Rates of return and capital aggregation using alternative rental prices. In *Technology and capital formation*, ed. D. Jorgenson and R. Landau. Cambridge, MA: MIT Press.
Hicks, J. 1939. *Value and capital*. Oxford: Clarendon Press.
Hulten, C. 1990. The measurement of capital. In *Fifty years of economic measurement*, *National bureau of economic research studies in income and wealth*, ed. E. Berndt and J. Triplett, Vol. 54. Chicago: University of Chicago Press.
Hulten, C. 1996. Capital and wealth in the revised SNA. In *The new system of national accounts*, ed. J. Kendrick. New York: Kluwer Academic Publishers.
Hulten, C., and F. Wykoff. 1981a. The estimation of economic depreciation using vintage asset prices. *Journal of Econometrics* 15: 367–396.
Hulten, C., and F. Wykoff. 1981b. The measurement of economic depreciation. In *Depreciation, inflation and the taxation of income from capital*, ed. C. Hulten. Washington, DC: Urban Institute Press.
Hulten, C., and F. Wykoff. 1996. Issues in the measurement of economic depreciation: Introductory remarks. *Economic Inquiry* 34: 10–23.
Jorgenson, D. 1989. Capital as a factor of production. In *Technology and capital formation*, ed. D. Jorgenson and R. Landau. Cambridge, MA: MIT Press.
Jorgenson, D. 1996. Empirical studies of depreciation. *Economic Inquiry* 34: 24–42.
Jorgenson, D., and Z. Griliches. 1967. The explanation of productivity change. *Review of Economic Studies* 34: 249–283.
Jorgenson, D. and Griliches, Z. 1972. Issues in growth accounting: A reply to Edward F. Denison. *Survey of Current Business* 52(5), Part II, 65–94.
Matheson, E. 1910. *Depreciation of factories, mines and industrial undertakings and their valuations*. 4th ed. London: Spon.
Nadiri, M., and I. Prucha. 1996. Estimation of the depreciation rate of physical and R and D capital in the U.-S. total manufacturing sector. *Economic Inquiry* 34: 43–56.
Oliner, S. 1996. New evidence on the retirement and depreciation of machine tools. *Economic Inquiry* 34: 57–77.
Pakes, A., and Z. Griliches. 1984. Estimating distributed lags in short panels with an application to the specification of depreciation patterns and capital stock constructs. *Review of Economic Studies* 51: 243–262.
Schreyer, P. 2006. Measuring multi-factor productivity when rates of return are exogenous. In *Price and productivity measurement*, ed. W. Diewert et al., Vol. 1 and 2. Vancouver: Trafford Press.
Triplett, J. 1996. Depreciation in production analysis and in income and wealth accounts: Resolution of an old debate. *Economic Inquiry* 34: 93–115.
Walras, L. 1874. *Elements of pure economics*, trans. W. Jaffé. Homewood: Richard D. Irwin, 1954.

Capital Perversity

Tatsuo Hatta

Neoclassical capital theory regards the interest rate as the market price of the composite factor ‘capital’. In this theory the interest rate is equal to the marginal product of capital, since the demand curve for capital is its marginal productivity schedule. Moreover, the theory assumes that capital obeys the law of diminishing returns just like any other factor, so that its demand curve is downward-sloping. In an economy where labour is the only primary factor and constant returns to scale prevail, this implies the following postulate: *as the interest rate falls, the capital – labour ratio increases*, which plays an important role in neoclassical growth theory and in comparative static analyses of interest rate determination.

Neoclassical capital theory also makes another closely related postulate: *as the interest rate falls, the output–labour ratio increases*. This postulate does not explicitly use the concept of aggregate capital. However, it too implies that ‘capital’ obeys the law of diminishing returns. For the output–labour ratio can be raised only when some input other than labour is increased behind the scenes, and in this economy capital is the only such input available.

Both postulates necessarily hold if output is produced by labour and a *single* capital input in a linear homogeneous production function, as in

the Clark–Ramsey production function. Cambridge economists, led by Robinson (1953–4, 1956), Champernowne (1953–4) and Sraffa (1960), criticized these postulates, however, for economies with heterogeneous capital goods, thus kindling the so-called Cambridge controversies in capital theory as surveyed by von Weizsäcker (1971), Harcourt (1972), Blaug (1974), and Burmeister (1980). Eventually, counter-examples that appeared in Pasinetti et al. (1966) showed irrefutably that both postulates can fail to hold in such economies. These paradoxical phenomena are called *capital perversities*. They showed very clearly that 'capital' is different from other factors in that diminishing returns do not hold for it even in contexts quite free of aggregation problems.

In order to examine the first postulate for economies with heterogeneous capital goods, one has to aggregate heterogeneous capital goods into a single dollar value of capital. Such a measure could well be specious, however, due to the index number problem involved in the aggregation, the interest rate affecting the prices of capital goods with different gestation periods differently. Since the second postulate does not depend on a particular aggregate measure of capital, it may appear a more robust characterization of diminishing returns from roundabout processes than the first. In fact, the following proposition due to Burmeister and Dobell (1970, Corollary 7.2) implies that the two postulates are equivalent once a proper price index is chosen for evaluating capital.

> Suppose than an exogenous increase in interest rate shifts one stationary-state production equilibrium to another. Then, as long as the interest rate is positive, the ratio of output to labour moves in the same direction as the ratio of 'constant-price capital' to labour, where the 'constant-price capital' is the dollar value of the new capital input vector measured at the initial input price vector.

For this reason we will examine only the failure of the second postulate to hold.

Reswitching

Capital perversity was demonstrated via examples of the so-called reswitching phenomenon; the simplest and most illuminating is Samuelson's (1966). He assumes that output this year is produced by applying labour inputs in three preceding years according to the following production function:

$$Y = y(x_1, x_2, x_3), \tag{1}$$

where Y is this year's output level and x_1, x_2, and x_3 are labour inputs one, two and three years ago, respectively. Let p_t be the present value of the wage rate t periods prior to the production year. Producers chose the cost-minimizing input vector (x_1, x_2, x_3) for the given output level under the input price vector (p_1, p_2, p_3). Samuelson also assumes free entry, so that maximized profit is zero.

Now consider a steady-state economy where Y is produced every year and prices are constant. Then we have

$$p_t = w \cdot r^t, \tag{2}$$

where w is the (constant) wage rate and r is 1 plus the interest rate. Input and output variables for each year may be shown as in Table 1. Each column shows the total amount of labour L applied in the entire production process that year as

$$L = x_1 + x_2 + x_3. \tag{3}$$

As the macroeconomist sees it, this economy as a whole produces Y every year by applying capital inputs in the form of goods-in-process and an amount L of labour.

Samuelson considered the case where the technology (Eq. 1) consists of only two techniques α and β: α's input vector (x_1, x_2, x_3) for producing a unit output is (0, 7, 0) and β's (6, 0, 2). He showed

Capital Perversity, Table 1

1986	1985	1984	1983	1982	1981	1980
			y	x_1	x_2	x_3
		y	x_1	x_2	x_3	
	y	x_1	x_2	x_3		
y	x_1	x_2	x_3			

that β minimizes cost when the interest rate lies between 50 and 100 per cent year, while α does so otherwise. As the interest rate increases from zero, therefore, the cost-minimizing technique switches first from α to β, and then back to α. This phenomenon, that as the interest rate increases, a once-abandoned technique becomes re-employed, is called *the reswitching of techniques*. It is obvious that when it happens capital perversity necessarily occurs. In Samuelson's case, for example, when the interest rate is increased past 100 per cent, technique β with $Y/L = 1/8$ is switched to α with 1/7, falsifying the second postulate. It can readily be shown that at this switching interest rate the first postulate also fails, even after the index number problem is removed.

What Causes Perversity?

Examples of reswitching had to be given for economies with discrete technologies, since it occurs with probability zero in a smoothly substitutable production function. But neither reswitching nor a discrete technology is necessary for perversity itself. Indeed, Hatta (1976) constructed an example of a smoothly substitutable and linear homogeneous function of type (Eq. 1) that behaves perversely.

To see how this might work, consider a generalized version of (Eq. 1):

$$Y = y(x_1, x_2, \ldots, x_n), \tag{4}$$

where y is quasi-concave, linear homogeneous, and differentiable. Then we have the following proposition due to Hatta (1976), which was independently hinted at by Solow (1975, p. 52):

For capital perversity to occur in Eq. 4 it must have at least one complementary input pair. (A) Equivalently, if all input pairs in Eq. 4 are (Hicksian) substitutes, perversity cannot occur.

According to a standard Hicksian demand rule (1946, ch. 3), (net) complementarity among inputs can occur in Eq. 4 only if n is greater than 2. Thus Proposition (A) implies that for perversity to occur in Eq. 4, n must be greater than 2. When $n = 2$, on the other hand, the economy has only one capital good, i.e., the one produced by the labour input applied in the previous year. Proposition (A) therefore implies that:

Heterogeneity of capital is necessary for perversity in Eq. 4. (B)

We now prove (A) for the case n = 3. The cost-minimizing input vector for output level Y under the input price vector (p1, p2, p3) is given by the following set of input demand functions:

$$x_s = a_s(p_1, p_2, p_3, Y) \quad s = 1, 2, 3.$$

We assume that the interest rate is positive, i.e.,

$$r > 1. \tag{5}$$

Noting Eq. 2 and the zero-degree homogeneity of a_s in the prices, the following must hold when cost is minimized:

$$x_s = a_s\left(1, r, r^2, Y\right) \quad s = 1, 2, 3.$$

In view of Eq. 3, therefore, the total labour movement requirement in this stationary economy is

$$L = a_1\left(1, r, r^2, Y\right) + a_2\left(1, r, r^2, Y\right) + a_3\left(1, r, r^2, Y\right)$$

By definition perversity occurs if the L necessary to produce a constant Y every year is lowered when the interest rate is raised, i.e., if

$$\partial L/\partial r < 0. \tag{6}$$

Carrying out this differentiation, we obtain

$$\frac{\partial L}{\partial r} = (r-1)a_{12} + 2\left(r^2 - 1\right)a_3 + \left(r^2 - r\right)a_{23} \tag{7}$$

where

$$a_{st} \equiv \partial a_s/\partial p_t.$$

This and Eq. 5 imply that $\partial L/\partial r$ is positive if all a_{st}'s (i.e. Hicksian cross-substitution terms) are positive. This in turn implies that for perversity

to occur, there must be at least one complementary input pair. Q.E.D.

For general n, Proposition (Eq. 6) is proved similarly, since Eq. 7 generalizes to

$$r \cdot \frac{\partial L}{\partial r} = \sum_{s=1}^{n-1} \sum_{t=s+1}^{n} (t-s)(p_t - p_s) \cdot a_{st}.$$

Now look at Samuelson's example in the light of (A). Assume that for given Y and given prices, β is cost-minimizing. Now let p_1 increase, keeping p_2 and p_3 constant. Eventually this will make β more costly than α, so α will be employed. But α uses less x_3 than β in order to produce the same output, so the rise in p_1 has caused a reduction in x_3, i.e. pair (1, 3) is complementary. Thus Samuelson's discrete model is consistent with our Proposition (A), obtained for the neoclassical production function.

Hence perversity is simply one of the many paradoxes caused by complementarity. The reason why the Clark–Ramsey production function always behaves well is now clear: it has only two inputs, which must be substitutes.

Why Complementarity?

Why does complementarity cause perversity? Note first that when $n = 3$ perversity cannot occur if either the input pair (1, 2) or the pair (2, 3) is complementary. Indeed, when $n = 3$ the following stronger version of (A) holds: For perversity to occur in Eq. 1, a_{13} must be negative, i.e., the specific input pair (1, 3) must be complementary. (C)

Just as a complementary pair of consumption goods can be regarded as a composite good, a complementary pair of inputs (e.g. truck and garage) may be treated as a composite input. When a neighbouring input pair is complementary in the production function (Eq. 1), therefore, that function can be regarded as containing just two inputs: one (composite) labour and a (composite) capital. For example, when (1, 2) is complementary, the pair (1, 2) can be regarded as composite labour. In such cases the production function is essentially of Clark–Ramsey form and so behaves well.

When (1, 3) is complementary, on the other hand, the technology's two (composite) inputs (1, 3) and 2 cannot be ranked in terms of their gestation periods. The two inputs can interchange the roles of capital and labour for different levels of interest rate, which explains why perversity can occur in this situation. Observe that (1, 3) is also complementary in Samuelson's model with a discrete technology, and the above explanation is applicable to his model.

Proposition (C) can be extended in various ways to the case where $n > 3$. For example, perversity never occurs if the structure of complementarity is such that the n inputs can be classified into one composite labour and one composite capital. Thus perversity occurs only if complementarity creates a composite input that cannot be unequivocally ranked with another (composite) input *vis-à-vis* their gestation periods. As Hatta (1976) argues, Bruno et al.'s (1966) non-reswitching condition can be interpreted in this spirit.

The proof of (C) is straightforward. Noting that $a_{11} + r\, a_{12} + r^2\, a_{13} = 0$ and $a_{31} + r\, a_{32} + r^2\, a_{33} = 0$, from the homogeneity property of the input demand functions in prices, we can rewrite (Eq. 7) as:

$$\frac{r}{w} \cdot \frac{\partial L}{\partial r} = (1-r) \cdot a_{11} + \left(r^3 - r\right) \cdot a_{13} + \left(r^3 - r^4\right) \cdot a_{33}.$$

This implies that a_{13} must be negative if perversity occurs, since r is greater than 1 and a_{11} and a_{33} are negative from the Hicksian demand rule. Thus (C) is proved.

Conclusion

To construct models of growth and the interest rate in an economy with heterogeneous capital– good inputs, the concept of 'capital' is not at all necessary: microeconomic production functions can be specified directly in terms of the physical units of those inputs. The main focus of the Cambridge controversies in capital theory was rather on the question of how well the simple Clark–Ramsey production function can approximate the

qualitative properties of a production economy with heterogeneous capital–good inputs.

It was established through these controversies that the monotonic relationship between output–labour ratio and interest rate, a basic property of the Clark–Ramsey production function, fails to hold in a world of heterogeneous capital inputs. Since this relation has nothing to do with the index number problem, the fact that it breaks down in a general model clearly contradicted that part of neoclassical capital theory which was based upon the Clark–Ramsey production function. This was a genuinely new finding that came out of the capital controversies. As we have seen, however, it is fully explicable within neoclassical theory, being no more (and no less) than one of the many intractable problems caused by the presence of complementarity.

See Also

- ▶ Capital Theory (Paradoxes)
- ▶ Reverse Capital Deepening

Bibliography

Blaug, M. 1974. *The Cambridge revolution: Success or failure?* London: The Institute of Economic Affairs.

Bruno, M., E. Burmeister, and E. Sheshinski. 1966. The nature and implications of the reswitching of techniques. *Quarterly Journal of Economics* 80: 526–553.

Burmeister, E. 1980. *Capital theory and dynamics*. Cambridge: Cambridge University Press.

Burmeister, E., and R. Dobell. 1970. *Mathematical theories of economic growth*. New York: Macmillan.

Champernowne, D.G. 1953–4. The production function and the theory of capital: A comment. *Review of Economic Studies* 21: 112–135.

Harcourt, G.C. 1972. *Some Cambridge controversies in the theory of capital*. London: Cambridge University Press.

Hatta, T. 1976. The paradox in capital theory and complementarity of inputs. *Review of Economic Studies* 43: 127–142.

Hicks, J. 1946. *Value and capital*, 2nd ed. London: Oxford University Press.

Pasinetti, L.L., et al. 1966. Paradoxes in capital theory: A symposium. *Quarterly Journal of Economics* 80: 503–583.

Robinson, J. 1953–4. The production function and the theory of capital. *Review of Economic Studies* 21: 81–106.

Robinson, J. 1956. *The accumulation of capital*. London: Macmillan.

Samuelson, P.A. 1966. A summing up. *Quarterly Journal of Economics* 80: 568–583.

Solow, R. 1975. Brief comments. *Quarterly Journal of Economics* 89: 48–52.

Sraffa, P. 1960. *Production of commodities by means of commodities prelude to a critique of economic theory.* Cambridge: Cambridge University Press.

von Weizsäcker, C.C. 1971. *Steady state capital theory.* Berlin: Springer.

Capital Theory

Robert A. Becker

Abstract

Capital theory examines the special role played by time in resource allocation studies. The determination of the interest rate and functional distribution of income as well as how rational agents invest are analysed within single- and multi-sector general equilibrium frameworks. Here, agents exercise perfect foresight over alternative consumption and capital accumulation programs. Efficient programs are characterized. Representative and multi-agent infinitely lived households are studied. Equivalence principles link the equilibrium programs and optimal paths. Heterogeneous agent models with borrowing constraints are reviewed. A behavioural model of intertemporal choice is also compared to its constant discounting counterpart.

Keywords

Aggregate capital; Aggregation; Allais paradox; Altruism; Arbitrage; Arbitrage pricing theory; Balanced growth; Behavioural economics; Cambridge controversies; Capital accumulation; Capital deepening; Capital theory; Capital value; Cobb–Douglas functions; Commitment; Comparative dynamics; Continuous and discrete time models; Deterministic models; Discounting; Dynamic non-substitution theorems; Dynamic

programming; Economic growth; Efficient allocation; Elasticity of substitution; Epstein–Hynes utility functions; Equivalence principle; Euler equations; Exhaustible resources; Existence of general equilibrium; Expected utility; Functional distribution of income; Futures markets; General equilibrium; Hahn problem; Hotelling, H.; Hyperbolic discounting; Impatience; Incomplete markets; Infinite horizons; Interest rates; Interest rate determination; Intertemporal choices; Intertemporal utility functions; Investment criteria; Fisher, I.; Kuhn–Tucker conditions; Long run and short run; Markov perfect equilibrium; Multi-capital goods models; Named goods; Neoclassical capital theory; No-arbitrage conditions; Non-classical production functions; Optimal growth; Orthodox vision of capital theory; Perfect foresight; Portfolio equilibrium; Present value investment criteria; Probability; Quasi-geometric discounting; Rae, J.; Ramsey, F. P.; Ramsey model; Rational expectations; Rationality; Recursive utility functions; Reduced form model; Renewable resources; Representative agent; Risk; Saving and investment; Shadow pricing; Spot markets; Sraffa, P.; Stationary state; Stylized facts; Technical change; Time; Time consistency; Time preference; Transversality condition; Turnpike theorems; Uncertainty; von Neumann, John

JEL classifications
E22

Introduction

Capital theory examines the special role played by time in resource allocation studies. The determination of the rate of interest and the functional distribution of income are considered along with the development of criteria for evaluating investment decisions. Contemporary capital theory focuses on the intertemporal choices undertaken by rational actors within a general equilibrium setting where all prices and allocations are determined by market clearing. The central role played by time is that producing goods and services to supply future consumption requires withdrawing some output from current consumption in order to create the produced means of production, or capital goods, which enable future production to be undertaken in conjunction with other factors such as labour and land. That agents seek to make their investment decisions rationally is taken as a fundamental premise of capital theoretic models. The rationality hypothesis is implemented by assuming that agents maximize a utility function over paths of future consumption and that producers maximize the present discounted value of their profits. A specification of the degree of foresight must be postulated together with an assumption on which spot and futures markets are open for trade. Consumption and investment decisions are realized in a market equilibrium.

Dated Commodities and Prices

The classical general equilibrium model developed over the last half of the 20th century by Arrow, Debreu, McKenzie and their followers was sufficiently abstract that it could model any number of different economic activities by the device of *named goods*: a commodity was specified by its physical characteristics, date of availability, contingent events upon which its availability depended, as well as its location. For example, a consumption good available now was differentiated from the same physical commodity available at a different date even if the location or contingent events were the same at both dates. Capital theoretic models focused on the pure role of time assume certainty (no contingent events) and the same location. The simplest models assume that there is just one consumption good and that its characteristics are the same at each point of time. Only the date of its availability differentiates goods. These are the *deterministic models*. Agents are supposed to exercise perfect foresight over the paths of all relevant variables in this case. Other models treat both time and uncertainty by way of dated goods and contingent events. Rational expectations about the future

probability distributions of variables are assumed to describe agents' behaviour. The basic principles and issues in capital theory are most easily reviewed in the deterministic setting with risk and uncertainty treated as a non-trivial extension of the basic theory.

The classical general equilibrium model assumes a finite number of commodities. In the deterministic intertemporal setting this means there are a finite number of dated commodities. Consumers have a finite planning horizon; time unfolds in discrete periods, $t = 1, 2, \ldots, T$. A finite number of goods are available at each date, indexed by $i = 1, 2, \ldots, N$. This makes for NT commodities. Consumers' preferences are defined over a commodity space contained in an NT-dimensional Euclidean space. Similarly, producers' technology sets were defined in the same commodity space. Competitive prices are established through a market mechanism on the presupposition that markets operate for all NT commodities. The classic existence of equilibrium and welfare theorems apply under appropriate assumptions on the consumption and production sectors as well as the relations between them. This formal connection between intertemporal and atemporal static general equilibrium theory offers little that is new or special to capital theory. It is the recognition that time places restrictions on preferences and technologies that specialize the abstract Walrasian model to the type more suited to answering capital theoretic questions about interest rate determination and the corresponding division of the model's output among its participating consumers and resource owners.

The distinguishing feature of capital theoretic models is their focus on infinite horizon decision problems. The motivation for this lies in the open-ended nature of the economic problem. Economies do not have foreseeable ends and the problem of saving and investing for future consumption seemingly goes on for ever, even though all the decision makers know that our planet's time is limited. But that terminal date is so far in the future that we might as well act today as if an infinite horizon is a good approximation to a very long but finite horizon. The theoretical advantage of the infinite horizon is that it allows us to draw a sharp formal distinction between the short and the long runs. The short run represents the transitional time that model solutions follow, whereas the long run constitutes the solutions' properties as time runs towards infinity. The classical focus on the stationary state, or 'long period', presumes there is a long run and that the economy evolves towards it.

Frank Ramsey (1928) modelled infinite horizons in a seminal article on optimal growth. He argued that discounting by the planner was ethically indefensible. Ramsey's modern followers from Paul Samuelson to the present day have studied both undiscounted and discounted models. Von Neumann's (1937) celebrated model of capital accumulation at a maximum balanced growth rate implicitly assumed an infinite horizon. A balanced program occurs when each type of capital good grows from one period to the next at the same constant rate. By focusing attention on balanced growth paths, it would seem reasonable that von Neumann understood those programs might correspond to that model economy's long-run position. The infinite horizon assumption has a long tradition in capital theory and finance (for example, the consol bonds issued by the United Kingdom; see Goetzmann and Rouwenhorst, 2005, for other examples).

This article concentrates entirely on the discounted case and its connection to general competitive analysis. The primary focus is taken to be the one-sector discounted Ramsey model. Capital theory is viewed as a branch of general equilibrium theory. The masterful surveys by McKenzie (1986, 1987) lay out the undiscounted as well as discounted models for many capital goods and multiple sectors in great generality. His surveys also provide details on how those models can evolve over time (the so-called turnpike theorems) as well as general comparative dynamics results.

Ramsey (1928) formulated his seminal model in continuous time. The models presented here are cast in discrete time with periods $t = 1, 2, \ldots$. This turns out to have some technical advantages over continuous time modelling as well as expositional advantages as economic concepts are more readily grasped by readers unschooled in

the calculus of variations and its modern development, optimal control theory.

Neoclassical Capital Theory: The One-Sector Model

The Discounted Ramsey Optimal Growth Model

Neoclassical capital theory is illustrated by the properties exhibited in the discrete time one-sector discounted Ramsey optimal growth model (Ramsey, 1928). This model encapsulates the fundamental consumption–investment trade-offs that a decision maker considers when choosing a consumption plan over time to achieve a maximum lifetime utility. The model is simplified in many ways. There is a single decision maker, or planner, acting over an infinite horizon. There is no uncertainty or shocks that would make output available in the future look like a random variable when viewed from the present. The model examines an aggregated economy. There is a single all-purpose consumption good produced using capital goods (carried over from the previous period) and fixed labour. The capital and consumption goods available at each time are physically identical and can be costlessly converted from consumption to capital (and vice versa) at a one-to-one rate. The planner decides how much to consume in the current period and how much to save for next period's production. Capital depreciates entirely within the period. It is *circulating* as it is used up within the production period. Extensions to include durable capital that depreciates at a fixed rate are straightforward. The planner's exogenously given initial stock of capital produces goods available in the first period. The planner obtains utility from consumption at each time and maximizes the discounted sum of future utilities. The discount factor on future utility is a given constant.

The planner's intertemporal optimization problem is:

$$\sup \sum_{t=1}^{\infty} \delta^{t-1} u(c_t) \text{ by choice of } \{c_t, k_{t-1}\}_{t=1}^{\infty}, \quad (1)$$

subject to:

$$c_t + k_t \le f(k_{t-1}) \text{ for } t = 1, 2, \ldots; \; c_t \ge 0, \quad k_{t-1} \ge 0 \text{ all } t; \; k_0 \le k, \quad \text{where } k > 0 \text{ is given.} \quad (2)$$

Feasible programs are sequences $\{c_t, k_{t-1}\}_{t=1}^{\infty}$ which satisfy (2). Assume $u : [0, \infty) \to [0, \infty)$ is strictly concave, increasing, twice continuously differentiable, $u(0) = 0$, and satisfies the Inada condition: $\lim_{c \to 0+} u'(c) = \infty$. The production function $f : [0, \infty) \to [0, \infty)$ is strictly concave, increasing, twice continuously differentiable, $f(0) = 0$, satisfies $\lim_{k \to 0+} f'(k) = \infty$, and $\lim_{k \to \infty} f'(k) < 1$ (also called Inada conditions). There is a maximum sustainable stock, $b > 0$, with $f(b) = b$ and $0 < k < b$. The discount factor, δ, satisfies $0 < \delta < 1; \delta = 1/(1 + \tilde{n})$, where $\tilde{n} > 0$ is the pure rate of time preference (or rate of impatience). There is a unique optimal program, $\{\overline{c}_t, \overline{k}_{t-1}\}_{t=1}^{\infty}$. Its discounted utility sums, $\sum_{t=1}^{\infty} \delta^{t-1} u(\overline{c}_t) < \infty$. The optimal growth problem has a *time consistency property*: The optimal sequence $\{\overline{c}_t, \overline{k}_{t-1}\}_{t=1}^{\infty}$ has the property that $\{\overline{c}_{t+\tau}, \overline{k}_{t-1+\tau}\}_{t=1}^{\infty}$ solves the optimization problem with objective starting at time τ, $\sum_{t=1}^{\infty} \delta^{t-1+\tau} u(c_{t+\tau})$, subject to $c_{t+\tau} + k_{t+\tau} \le f(k_{t-1+\tau})$ for $t = 1, 2, \ldots$ and $k_\tau = k$. Calender time is irrelevant: if the planner's objective is moved forward τ periods and the initial capital stock is maintained at the new starting time, then the optimal capital and consumption sequence are identical to the ones initiated at time $\tau = 0$. The reason for this is $\sum_{t=1}^{\infty} \delta^{t-1+\tau} u(c_{t+\tau}) = \delta^{\tau} \sum_{t=1}^{\infty} \delta^{t-1} u(c_{t+\tau})$, which is multiple of (1) and the set of feasible programs is unchanged. Hence, the optimal solution is unchanged from the same initial condition even though time has simply been reset to start at τ.

The optimal program satisfies $(\overline{c}_t, \overline{k}_{t-1}) > 0$ for each t. The Kuhn–Tucker necessary conditions for an optimum, known as the *Euler*, or *no-arbitrage conditions*, are:

$$\delta f'(\overline{k}_t) u'(\overline{c}_{t+1}) = u'(\overline{c}_t), \quad \text{for each } t. \quad (3)$$

If the planner's horizon is a finite period, T, then (3) and the complementary slackness

condition $\delta^{T-1}u'(\overline{c}_T)\overline{k}_T = 0$ obtain. The latter condition states capital's terminal value is zero. For the infinite horizon case of interest, it is natural to conjecture the *transversality condition* holds as a necessary condition for optimality:

$$\lim_{T\to\infty} \delta^{T-1}u'(\overline{c}_T)\overline{k}_T = 0. \tag{4}$$

This condition's necessity can be formally demonstrated in many problems. The conditions (3) and (4) are also sufficient conditions for optimality under the maintained hypotheses governing the concavity of the single period return function, u, and the production function, f.

Equation (3) expresses the unprofitability of the *one-period reversed arbitrages* developed below. An *arbitrage* represents a feasible change in the optimal path. Reversed arbitrages perturb the optimum for finitely many consecutive periods. Unreversed arbitrages change the optimal path permanently from some given time on to infinity. A necessary condition for an optimal path is that no arbitrage increase the discounted sum of future utilities above the optimal discounted utility. The necessity of the transversality condition can be interpreted as a type of no-arbitrage condition for *unreversed arbitrages* which never return to the optimal path.

Suppose that the consumption and capital sequences $\left(\overline{c}_t, \overline{k}_{t-1}\right) > 0$ (for each t) are optimal for the given initial capital stock. Then, the planner cannot increase utility by undertaking the following activity: at time t marginally increase the capital stock to be carried to time $t + 1$. This costs the planner $u'(\overline{c}_t)$ utils on the margin. Now invest this extra capital to obtain $f'\left(\overline{k}_t\right)$ additional units of goods in period $t + 1$ from the production sector. Convert this additional income into consumption at $t + 1$ worth $u'(\overline{c}_{t+1})$ utils on the margin. This implies the marginal benefit of this incremental investment measured at $t + 1$ is $f'\left(\overline{k}_t\right)$ $u'(\overline{c}_{t+1})$. Now discount this by the utility discount factor δ to place the marginal benefit at time $t + 1$ and marginal cost at time t in comparable utility units. The marginal benefit cannot exceed the marginal cost along an optimal solution to the household's problem. This is formally expressed by the inequality $\delta f'\left(\overline{k}_t\right)u'(\overline{c}_{t+1}) \leq u'(\overline{c}_t)$, for each t. Since the capital stock at time t is positive, then this arbitrage calculation can be repeated for an increase in consumption at time t paid for by lower consumption at time $t + 1$. In this case, the inequality is reversed and (3) holds.

This model has one special solution: it is the stationary optimal program (c^*, k^*), with $c^* = f(k^*) - k^*$ and $\delta f'(k^*) = 1$. By concavity of f, this program has the property that k^* solves the problem $\max_{k\geq 0}[\delta f(k) - k]$. This is a form of the dynamic non-substitution theorem: the stationary optimal capital stock is independent of the planner's felicity function and depends only on technology and the planner's discount factor. The equation $\delta f'(k^*) = 1$ is also the Euler equation for the program $c_t^* \equiv c^*$ and $k_{t-1}^* \equiv k^*$ for each $t \geq 1$. That is, if the initial capital stock is k^*, then it is optimal to maintain that capital stock for ever. The program $\left\{c_t^*, k_{t-1}^*\right\}_{t=1}^{\infty}$ is constant, or stationary, over time. Hence the name: the stationary optimal program (also called the *steady state*). In the case $\delta = 1$ the steady state maximizes stationary consumption over all feasible stationary consumption levels (it is the *optimal stationary consumption path*) and is called the *golden-rule consumption level* while the corresponding stationary capital stock is the *golden-rule capital stock*. For the discounted case, $0 < \delta < 1$, the steady states are also known as the *modified golden-rule consumption* and *capital stock*.

The optimal path of the infinite horizon problem with initial stocks $k \neq k^*$ converges monotonically to the stationary optimal program (c^*, k^*), with $c^* = f(k^*) - k^*$ and $\delta f'(k^*) = 1$. For example, if $0 < k < k^*$, then the optimal capital sequence, $\left\{\overline{k}_{t-1}\right\}_{t=1}^{\infty} \nearrow k^*$. Moreover, paths *do not cross*: if $0 < k < k' < k^*$, then $\overline{k}_t < \overline{k}'_t$, where $\left\{\overline{k}'_{t-1}\right\}_{t=1}^{\infty}$ is optimal from initial stocks, k'. The convergence of the optimal path implies it is bounded, and the transversality condition holds as a necessary condition for optimality in this model. Conversely, a feasible program satisfying the Euler equations and transversality condition is an optimal program. The convergence property of the optimal capital sequences is also known as the *turnpike theorem*: the optimal

C

capital sequence from any initial starting stock converges to the modified golden-rule capital stock. The corresponding consumption sequences likewise converge (monotonically) to the golden-rule consumption level. The turnpike theorem's conclusion suggests that there is a distinction between the economy's *long-run steady state* and the *short-run transitional dynamics* that describe how the economy approaches that stationary optimal program. One consequence of the turnpike theorem is that optimal programs spend infinitely many periods in any neighborhood of the steady state. In that sense, the steady state is a good approximation for the transitional dynamics over long periods of time. The choice of the analyst lies in determining how small that neighbourhood is, and hence how many periods the economy is not 'sufficiently close' to the model's long-run solution.

The Canonical Example

The *logarithmic utility, Cobb–Douglas production economy* is an important example of Ramsey's optimal growth problem. Many writers refer to it as the *canonical example* of the one-sector model since its solution is explicitly found. The planner's single period utility function is $u(c_t) = \ln c_t$ and the production function has the Cobb–Douglas form $f(x) = x^{\rho}$ where $0 < \rho < 1$ is a technology parameter (it is capital's constant share of total income in a competitive equilibrium setting). The Ramsey optimal growth problem for this specification (and no depreciation) can be solved explicitly by a variety of techniques (see Becker and Boyd, 1997, for one such approach based on symmetry techniques). The solution is described by the *consumption policy functiong*$(k) = (1 - \delta_{\rho})k^{\rho}$ and the *capital policy functionh*$(k) = \delta\rho k^{\rho}$. At each date, the policy functions tell the decision maker how much to consume and how much to save given the current level of the capital stock, k. The optimal capital and consumption sequences are given by iterating the policy functions. Carrying out that iteration for example leads to the explicit solution for the capital sequence:

$$x_t(k) = \left(\delta_{\rho}\right)^{\rho^{t-1}+\cdots+1} k^{\rho^t} \tag{5}$$

The capital and consumption policy functions in this example have constant marginal propensities to save and consume, respectively. Solow's (1956) growth model postulated savings and consumption functions of this type within a one-sector framework with a Cobb–Douglas production function in order to model the process of economic growth. Solow also assumed exogenous technological progress in the form of labour augmenting technical change, whereby each worker becomes more productive at an exponentially growing rate. Solow aimed his model at describing stylized facts of economic growth. The model was not formally set up to reflect microeconomic based optimizing behaviour at the level of individual consumption–saving decisions. The canonical version of Ramsey's discounted model provides such a microfoundation for Solow's descriptive theory in case there is no exogenous technical progress.

Let $k_t = x_t(k)$. The policy functions satisfy the no arbitrage condition. Let $c_t = (1 - \delta\rho)k_{t-1}^{\rho}$ and $c_{t+1} = (1 - \delta\rho)k_t^{\rho}$, where k_t is the capital stock at time t. The no arbitrage condition is:

$$\frac{c_t}{\delta c_{t-1}} = \frac{(1 - \delta\rho)k_t^{\rho}}{\delta(1 - \delta\rho)k_{t-1}^{\rho}} = \rho k_t^{\rho-1} = f'(k_t).$$

This solution can also be shown to satisfy the transversality condition, which takes the form here:

$$\lim_{t\to\infty} \frac{\rho k_{t-1}^{\rho-1}\delta^{t-1}}{c_t} = 0.$$

Therefore, the policy functions tell us how to find the optimal solution to this optimal growth problem. The optimal policy functions have the time consistency property as well.

The qualitative features of the optimal solution also follow from the policy functions. The most important observation is that the optimal capital sequence is monotonic as can be shown by iterating the capital policy function. Notice that each

optimal path converges to the unique positive fixed point of the capital policy function, k^*, where $h(k^*) = k^*$, which implies that:

$$k^* = (\delta\rho)^{\frac{1}{1-\rho}}.$$

This is the model's modified golden-rule capital stock. If the positive initial capital is below the modified golden rule, then the economy accumulates capital and the sequence of optimal capital stocks increases and converges to the modified golden-rule capital stock. Similarly, the optimal capital stocks decrease and converge to the modified golden rule when the starting stock is larger than the positive fixed point. If the initial capital happens to equal the modified golden-rule stocks, then it will be optimal to maintain those stocks in every period. Thus, the modified golden rule is a steady state of the dynamical system:

$$k_{t+1} = h(k_t) = \delta\rho k_t^{\rho}.$$

The corresponding consumption sequence is also monotonic since the consumption policy function is increasing in capital. The resulting consumption sequence converges to the *modified golden-rule consumption* level defined by:

$$c^* = (1 - \delta\rho)(k^*)^{\rho}.$$

The convergence of the optimal capital and consumption sequences illustrates the *turnpike theorem*. The monotonicity property for optimal capital sequences can also be viewed as a *non-crossing property*: if $k < k'$ are two different starting stocks, then $h(k) = k_1 < k_1' = h(k')$. Continuing in this way we see that, when two starting stocks are compared, the lower one always provides less capital than the higher one at any time along the optimal program.

The steady state's sensitivity to the discount factor is readily shown for $0 < \delta < 1$ for the general discounted one-sector model. Let $k^* = k^*(\delta)$ denote the steady state capital stock as a function of the discount factor. The condition $\delta f'(k^*(\delta)) = 1$ implies upon differentiation that $dk^*/d\delta > 0$. This *comparative steady state* result means that a more patient planner (there is a marginal increase in discount factor) produces a larger stationary optimal capital stock. Some writers on capital theory call this the *capital deepening* response to a change in the discount factor. The corresponding result for the consumption path $c^*(\delta) = f(k^*(\delta)) - k^*(\delta)$ states $dc^*/d\delta > 0$ as well. This is called *non-paradoxical consumption behaviour*. Note that this comparative steady state exercise does not compare the optimal program starting from k^* given the new discount factor to the optimal stationary plan k^* for the old discount factor. Comparative steady state exercises merely compare the steady states before and after a parameter change without evaluating the economy's transition path from one steady state to another.

Comparative dynamics results are available for the one-sector model which include studying the transition from one steady state to another in response to a parameter change. The planner considers all feasible plans in response to a change in one of the economy's deep taste or technology parameters. In particular, it is possible to compare the optimal programs before and after the parameter changes. For example, if the planner's discount factor increases (or, equivalently, the pure rate of time preference declines), then the planner becomes more patient. If the planner's discount factor increases from δ to δ', with $0 < \delta < \delta' < 1$, then the optimal capital paths starting from the same initial capital stock satisfy the conditions $\overline{k}'_t > \overline{k}_t$ for each time – there is a *generalized capital deepening response* because the economy's capital stock is increased at each time. Indeed, the discount factor's *initial impact* is to increase the first period's capital stocks at the expense of first period consumption since the initial capital stocks and first period output are unchanged after the discount factor increases. As the new consumption program converges monotonically to a larger modified golden-rule consumption level, $c^*(\delta')$, it follows that *eventually* (that is, in finite time) $\overline{c}_t(\delta') > \overline{c}_t(\delta)$ must obtain. These comparative dynamics results are easily verified for the canonical example with log utility and Cobb–Douglas production.

It is interesting to note that the monotonicity and non-crossing properties of the one-sector

model are robust. For example, the concavity of the production function can be relaxed while preserving these qualitative properties. The production function is *non-classical* provided there is an inflection point, $0 < k_I < b$ such that $f''(k) > 0$ for $k < k_I$ and $f''(k) < 0$ for $k > k_I$. Non-classical production functions can arise in fishery models when representing the production of a new generation of fish from the existing population. See Becker and Boyd (1997, Chap. 5) for details on the non-classical production extensions.

Generalizations of the one-sector model's turnpike property (the convergence of optimal capital sequences to the modified golden-rule stock) are also available for some multi-capital goods models, as found in McKenzie's surveys. The original turnpike theorem for many capital goods models was conjectured by Dorfman et al. (1958) in the von Neumann model framework without an explicit consumption criterion. Radner (1961) provides the first rigorous proof of a turnpike theorem for a von Neumann style model with a unique maximum balanced growth path and a finite planning horizon. Radner's theory evaluated alternative programs from a given initial vector of capital stocks according to a criterion based on the value of those stocks in the program's final time period. As with Dorfman, Samuelson and Solow's model, Radner's theorem did not apply to a Ramsey-style planner with an objective based on discounted utility. Radner's value loss technique for demonstrating the turnpike theorem did turn out to apply to undiscounted Ramsey models as well as some forms of the discounted model, as summarized in McKenzie's survey articles.

Another generalization focuses on the representation of the intertemporal utility function. Some recursive utility functions, which generalize the time consistency property of the time additive utility function, can be specified for concave production models while retaining the qualitative properties of optimal paths, such as capital monotonicity. The basic notion of a recursive utility function is illustrated below. The general theory of recursive utility functions is exposited by Becker and Boyd (1997).

Flexible time preference underlies many classic writings on capital theory – the agents discount factor depends on the underlying consumption stream. Recursive utility functions are one family of utilities that allow the steady state consumption stream to influence the corresponding discount factor. The brief development of recursive utility theory given here is grounded in a re-examination of the time consistency property of the planner's optimal choice in the one-sector discounted Ramsey model.

The discounted additive utility function, U, over infinite consumption streams $\mathbf{c} = \{c_1, c_2, \ldots\}$ is defined by the formula:

$$U(\mathbf{c}) = \sum_{t=1}^{\infty} \delta^{t-1} u(c_t)$$

where u is a bounded, strictly increasing, and strictly concave function on $[0, \infty)$ with $0 < \delta < 1$ as before. The time consistency property discussed above reflects the property that U is *recursive*: the behaviour embodied in this additive representation of utility has a self-referential property, that is, the behaviour of the planner over the infinite time horizon $t = 1, 2, \ldots$ is guided by the behaviour of that agent over the tail horizon $t = T, T + 1, T + 2, \ldots$ (for each T) hidden inside the original horizon. For this additive utility function, recursivity means the objective from time $T+1$ to $+\infty$ has the same form as the objective starting at time $T = 0$ (except for some time shifts in consumption dates). Formally, U may be rewritten as:

$$\sum_{t=1}^{\infty} \delta^{t-1} u(c_t) = \sum_{t=1}^{T} \delta^{t-1} u(c_t) + \delta^T \sum_{t=T+1}^{\infty} \delta^{t-1} u(c_{t+T}),$$

where the last sum gives the utility of the stream $\{c_{T+1}, c_{T+2}, \ldots\}$. The utility of the consumption stream $\mathbf{c}$ can be written as the function:

$$U(\mathbf{c}) = u(c_1) + \delta U(S\mathbf{c}),$$

where S is the *shift operator*: $S\mathbf{c} = \{c_2, c_3, \ldots\}$. Let the *projection operator*, π, be defined by the formula $\pi\mathbf{c} = c_1$. The general notion of a recursive

utility function is that the utility function U can be written in the form:

$$U(\mathbf{c}) = W(u(\pi\mathbf{c}), U(S\mathbf{c}))$$

for an appropriate real-valued function W defined on $[0, \infty) \times \mathcal{U}$, where $\mathcal{U}$ is the range of U. W is called the *aggregator function*. For the additive function, $W(c, y) = u(c) + \delta y$ for $y \in \mathcal{U}$. There are other examples of recursive utility functions. The Epstein–Hynes utility function developed below is generated by the *EH aggregator* $W(c, y) = (-1 + y) \exp(-v(c))$, where v is a strictly concave, increasing function of c with $v(0) > 0$.

The general theory of recursive utility functions provides a way to recover the utility function U from specification of the aggregator. Intuitively, U can be found by recursively substituting it into the equation $U(\mathbf{c}) = W(u(\pi\mathbf{c}), U(S\mathbf{c}))$. This substitution is performed by the *recursive operator* T_W defined by:

$$(T_W(U^0)(\mathbf{c}) = W(u(\pi\mathbf{c}), U^0(S\mathbf{c})),$$

where U^0 is considered the initial seed in this recursive substitution. For example, if $U^0 = 0$, the zero function that annihilates all consumption streams, then the $N^{th} - iterate$ of TW is:

$$(T_W^N 0) = W(c_1, W(c_2, \ldots, W(c_N, 0) \cdots)).$$

The recursive utility function is the unique fixed point of the operator T_W. The general theory provides conditions under which T_W has a unique fixed point and the successive iterates T_W^N converge to that fixed point independently of the choice of the initial seed function, U^0. Lucas and Stokey (1984) first proposed the specification of utility functions via aggregators and provided the basic theory of the recursion operator for bounded aggregators when consumption streams were elements of the set of all real-valued non-negative bounded sequences.

The basic ideas in recursive utility theory are readily illustrated for the case of the EH aggregator. This yields an example where the planner's utility function has flexible time preference and a recursive structure. A planner whose preferences over consumption streams is defined by the EH aggregator can be shown by recursive substitution to have the utility function U, which takes the form:

$$U(\mathbf{c}) = -\sum_{t=1}^{\infty} \exp\left(-\sum_{s=1}^{t} v(c_s)\right) \tag{6}$$

where $v : \mathbb{R}_+ \to \mathbb{R}_+$ is strictly concave, increasing, and satisfies $v(0) > 0$. Equation (6) is known as the *Epstein–Hynes (EH) utility function* after the continuous time analogue from Epstein and Hynes (1983); (6) was also studied in Epstein (1983). The EH utility from the consumption sequence's tail, $(c_{T+1}, c_{T+2}, \ldots)$, appears in the last term of the following expression breaking down the utility over the entire consumption path into segments for the first T periods and the subsequent periods:

$$\begin{aligned}
&-\sum_{t=1}^{\infty} \exp\left(-\sum_{s=1}^{t} v(c_s)\right) \\
&= -\sum_{t=1}^{T} \exp\left(-\sum_{s=1}^{t} v(c_s)\right) \\
&\quad + \exp\left(-\sum_{\tau=1}^{T} v(c_s)\right) \\
&\quad \times \left[-\sum_{t=T+1}^{\infty} \exp\left(-\sum_{\tau=T+1}^{t} v(c_s)\right)\right]
\end{aligned}$$

Hence, the utility of the tail of the program is just a time-shifted form of the utility of the original program – this is the identifying characteristic of a recursive utility function based on stationary preferences.

The steady state conditions for this economy are found by working out the no arbitrage conditions for the optimal growth problem which maximizes (6) subject to (2) and letting the consumption and capital sequences be constant sequences. Then the steady state conditions become:

$$f'(k^*) = 1/\exp(v(c^*)), \tag{7}$$

where k^* is the aggregate steady state capital stock. Since $\exp(v(0)) > 1$ and $c^* + k^* = f(k^*)$, one can solve (7) for a unique long-run capital and consumption level. The capital monotonicity property holds for the optimal solution to the problem of maximizing (6) subject to (2) when the neoclassical production function satisfies the concavity and Inada conditions for the discounted Ramsey model (see Becker and Boyd, 1997, Chap. 5, for a detailed proof and Beals and Koopmans, 1969, for the seminal article on recursive utility in optimal growth theory). In particular, if the initial capital stock is smaller than the steady state stock, then the economy's capital stock increases at each time and converges to the steady state; likewise, an initial capital stock above the steady state leads to a declining capital stock over time which converges to the steady state stock. The non-crossing property also obtains.

Equilibrium Equivalence Principles

The optimal growth model connects to the central questions of the determination of prices, including the rate of interest, and the functional distribution of income, by way of reinterpreting the optimal program as a competitive equilibrium for a fully specified dynamic general equilibrium model. This relationship is obtained by proving a version of the fundamental welfare theorems for this economy. The traditional welfare theorems based on finitely many goods must be adapted to the case of infinitely many dated commodities. There is more than one way to interpret the equilibrium model. The first interpretation is one with perfect foresight and a sequence of budget constraints, one for each time. Prices are reckoned in units of current consumption. The second interpretation links the neoclassical model with Irving Fisher's theory of interest rate determination and emphasizes his famous *separation principle*. The Fisherian equilibrium model is also one where agents act with perfect foresight.

At the core of either equilibrium model's interpretation is what Christopher Bliss (1975) called the *orthodox vision of capital theory*: an economy accumulating capital will generate rising wages and a falling rate of interest. Since capital increases over time, labour–capital complementarity implies workers are more productive and their wage rises. Diminishing returns set in and the rental rate falls as so many early writers on capital theory hypothesized in their verbal models. One of Ramsey's great contributions was to provide a consistent mathematical model of this story.

The PFCE Equivalence Principle

The competitive economy consists of an infinitely lived representative *household*, or *consumer sector*, and a *production sector*. The representative household's preferences coincide with the Ramsey style planner introduced above. The representative household is derived for an economy with a continuum of identical infinitely lived households whose preferences coincide with the Ramsey style planner. These households' preferences and endowments are identical. The total labour supply of all households has unit mass. In a *symmetric equilibrium* each household will take the same action given the same endowment, so it is sufficient to examine the decisions undertaken by a representative household who is also taken as supplying the economy's labour services to the production sector. The production sector's production function is the same as the one in the corresponding optimal growth model.

The representative consumer forecasts sequences of rental and wage rates to maximize lifetime utility subject to a sequence of budget constraints, one for each period. Formally, the household sector solves for given $\{r_t, w_t\}_{t=1}^{\infty}$ the problem:

$$\sup \sum_{t=1}^{\infty} \delta^{t-1} u(c_t)$$

by choice of the non-negative sequences $\{k_{t-1}, c_t\}_{t=1}^{\infty}$ subject to:

$$c_t + k_t = w_t + (1 + r_t)k_{t-1} \quad \text{for } t = 1, 2, \ldots \tag{8}$$

and $k_0 \leq k$. Here k is the initial capital stock (the same one as in the Ramsey optimal growth problem), r_t is the one-period rental rate on capital, and w_t is the wage rate earned by inelastically supplying one unit of labour in each time period. The prices r_t and w_t are reckoned in units of consumption available at time t.

The consumer's problem has a no arbitrage condition analogous to the one obtained in the optimal growth problem:

$$\delta(1 + r_t)u'(c_{t+1}) = u'(c_t) \quad \text{for each } t.$$

The transversality condition is necessary for equilibrium programs as defined below. The combination of the transversality condition and the no arbitrage equation is also sufficient for a consumption–capital sequence to solve the consumer's problem for a given profile of wages and rental factors.

Producers take the rental rate as given and solve the following myopic maximization problem for the production sector's capital demand at each time period:

$$\sup_{x \geq 0} f(x) - (1 + r_t)x.$$

Here, x denotes a level of aggregate capital; the profit maximizing solution is denoted k_{t-1}, the planned capital demand at time t. It only depends on the current rental rate, r_t. The problem's point input–point output structure reflects the absence of adjustment costs or other structural production lags and the fact that all forward-looking consumption–investment decisions reside in the household sector. The necessary and sufficient condition for a positive capital stock to solve the production sector's optimization problem at time t is:

$$f'(k_{t-1}) = 1 + r_t,$$

which uniquely determines k_{t-1} in terms of $1 + r_t$. The *total capital income* is $(1 + r_t)k_{t-1} = f'(k_{t-1})k_{t-1}$.

The *wage bill* is the residual 'profit' given by:

$$w_t = f(k_{t-1}) - (1 + r_t)k_{t-1}.$$

Notice that $w_t = f(k_{t-1}) - f'(k_{t-1})k_{t-1}$. In the Cobb–Douglas case with $f(k) = k^\rho$, then this economy labour's share of the total output or national product, k_{t-1}^ρ, is $1 - \rho$ and capital's share is ρ. The *total supply of goods in period t* is $f(k_{t-1})$ as a result of one-period profit maximization.

Sequences $\{1 + r_t, w_t, c_t, k_{t-1}\}_{t=1}^{\infty}$ constitute a *perfect foresight competitive equilibrium (PFCE)* provided that:

(PFCE-1) $\{c_t, k_{t-1}\}_{t=1}^{\infty}$ solve the consumer's problem given $\{1 + r_t, w_t\}_{t=1}^{\infty}$;
(PFCE-2) $f'(k_{t-1}) = 1 + r_t$, and
(PFCE-3) $w_t = f(k_{t-1}) - (1 + r_t)k_{t-1}$ for each time t.

These three conditions yield via Walras's Law the materials balance condition, $c_t + k_t = f(k_{t-1})$ for each t and $k_0 = k$.

The *equivalence principle* tells us that for this dynamic economy the *PFCE* allocation is the same as the Ramsey planner's solution. Hence, a *PFCE* allocation is an optimum and vice versa. The argument is the no arbitrage conditions for the equilibrium and optimal growth problems coincide, and the respective transversality conditions hold as necessary conditions in their respective problems. The sufficiency of these conditions is used to finish the proof.

A PFCE determines the *functional distribution of income* as the payments to each productive factor at each point in time. Labour receives its wage and capital is paid its capital income. The share of income received by each factor is a constant and time independent when production is Cobb–Douglas. The functional distribution of income at each time also yields the representative agent's personal income by adding the two source's income at each time. Multi-agent models differentiate the personal income an agent enjoys at each time from the corresponding functional distribution of income.

The Fisher Competitive Equilibrium Equivalence Principle

The capital theoretic foundation for the present value investment criterion is the *Fisher separation*

principle derived from Fisher's 'second approximation', which portrays the intertemporal consumption–investment decision of agents as a two-stage process. In the first stage, investment opportunities are exploited to realize a maximum value of initial wealth. The solution to the first-stage problem is found by maximizing the net present value over all feasible projects. Given competitive prices (and implicit discount rates), all agents whose intertemporal utility functions satisfy a mild non-satiation requirement will be led to choose the same wealth maximizing investment projects. In the second stage, those agents take their maximized wealth and access perfect capital markets to borrow and lend in order to obtain the most preferred lifetime consumption pattern.

The Fisher competitive equilibrium is the infinite horizon analogue of the Fisher separation principle. There is a single lifetime budget constraint; the savings–investment decision is separated from the consumption decision. Consumers maximize utility given their maximized wealth obtained as residual claimants to the production sector's discounted profit streams. Discounted profits are maximized within that sector. Letting $\{r_t\}$ be the sequence of interest rates and $q_t = \Pi_{\tau=1}^{t}(1+r_\tau)^{-1}$ the discounted price of time t consumption, define the profit function by $\pi(k,\{r_t\}) = \max\{\sum_{t=1}^{\infty} q_t[f(k_{t-1}) - (1+r_t)k_{t-1}] : k_0 = k\}$.

A sequence $\{r_t, c_t, k_t\}$ forms a Fisher competitive equilibrium (FCE) if:

(FCE-1) $\pi(k,\{r_t\}) = \max\{\sum_{t=1}^{\infty} q_t[f(k_{t-1}) - (1+r_t)k_{t-1}] : k_0 = k\}$;

(FCE-2) Consumers maximize $\sum_{t=1}^{\infty} \delta^{t-1}u(c_t)$ subject to the budget constraint $\sum_{t=1}^{\infty} q_t c_t = \pi(k,\{r_t\}) + k$;

(FCE-3) The market clearing condition $c_t = f(k_{t-1}) - k_{t-1}$ holds.

Once again, by matching first-order conditions and transversality conditions the sufficiency conditions for the agents' optimization problems imply that the allocation $\{c_t, k_t\}$ in a FCE $\{r_t, c_t, k_t\}$ is an optimum, and vice versa: given the optimal allocation $\{c_t, k_t\}$, there is a sequence of interest rates such that the triple $\{r_t, c_t, k_t\}$ forms a FCE. The result is the Fisher equivalence theorem.

The twin equivalence theorems for the PFCE and FCE models connect Ramsey's theory of optimal growth in an aggregate economy to Fisher's theory of consumption and investment in an intertemporal choice market model as well as to Solow's descriptive growth theory (the logarithmic utility, Cobb–Douglas production function example has a constant marginal propensity to save, as assumed in Solow's growth model). The qualitative properties of the optimal growth model carry over to the two formulations of dynamic competitive economies. In the case where the initial capital stocks are smaller than the modified golden-rule stocks, the capital monotonicity property of the optimal program implies that the consumption sequence increases, the sequence of wage rates is increasing, and the sequence of interest rates/ rental rates is decreasing. The orthodox vision of capital theory holds for the one-sector optimal growth model once the dynamic equilibrium is interpreted by way of the PFCE and FCE equivalence principles.

Many Agents

The equivalence principles for the discounted Ramsey model postulate a representative agent. The orthodox vision of capital theory carries over to some forms of neoclassical capital theory when many distinct agents replace the assumption of a representative infinitely lived household. The introduction of many distinct consumers raises interesting questions concerning the determination of equilibrium prices and the distribution of personal (and factor) income both in short and long runs.

Frank Ramsey's seminal contribution to optimal growth also addressed the long-run, or steady state, distribution in a competitive economy. He conjectured that, with households having different rates of impatience, the steady state equilibrium would have very unequal income and wealth distributions. The most patient household would enjoy the maximum sustainable consumption ('bliss' in his conception) and all other households would consume at a minimal level necessary to sustain their lives. This was not a particularly new idea at the time his paper was published. The

notion that time preference differences operating in a market economy might promote long-run differences in income and wealth can be found in the writings of such eminent economists as John Rae in 1834 and in several books by Irving Fisher beginning with his great work on the rate of interest first published in 1907. The Ramsey conjecture can be examined in two distinct neoclassical settings. The first deals with a natural extension of the optimal growth model to one of Pareto optimal growth. Agents are allowed to borrow and lend. The equilibrium version is analogous to the FCE set-up. Households have a single budget constraint expressed in present value terms. Here, long-run income distribution can be extreme if individuals have different discount factors – the relatively impatient ones receive NO income. The second formulation is one of temporary equilibrium where markets are incomplete – households are forbidden to borrow against their future labour income (each person's capital stock is constrained to be non-negative at each time) and face a sequence of budget constraints, as in the PFCE model. In this setting, the relatively impatient households consume their wage income and the most patient household consumes wage and capital income – a modern formulation of Ramsey's two-class society.

Pareto Optimal Growth with Many Agents

Suppose there are H households ($h = 1, 2, \ldots, H$) with one-period return functions u_h of the type met in the optimal growth setting. Let c_t^h denote agent h's consumption at time t and suppose that each agent's discount factor is the same $\delta = \delta h$ with $0 < \delta < 1$. Introduce *welfare weights* $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_H) \geq 0$ and $\sum_{h=1}^{H} \lambda_h = 1$. Given a weight vector λ, the Pareto optimal growth problem is to solve:

$$\sup \sum_{t=1}^{\infty} \sum_{h=1}^{H} \lambda_h \left[\delta^{t-1} u_h\left(c_t^h\right)\right] \tag{9}$$

$$\text{subject to } \left(\sum_{h=1}^{H} c_t^h\right) + k_t \leq f(k_{t-1}), t = 1, 2, \ldots,$$

$$c_t^h, k_{t-1} \geq 0, k_0 \leq k, h = 1, 2, \ldots$$

The planner seeks a path of consumption for each person and an aggregate capital path satisfying the constraints with the maximum weighted discounted future utility. This problem can be rewritten in an interesting manner.

Given a weight vector λ, define on $\mathbb{R}_+$ the real-valued function u_λ^* as the following program's optimal value function:

$$u_\lambda^*(c) = \sup\left\{\sum_{h=1}^{H} \lambda_h u_h\left(c^h\right) : \sum_{h=1}^{H} c^h = c, c^h \geq 0\right\}. \tag{10}$$

If u_h is a concave, continuous, increasing function on $[0,\infty)$, and twice continuously differentiable function on $(0,\infty)$, then u_λ^* is concave, increasing in c, and continuously differentiable. Note that the Inada condition $u_h'(0) = +\infty$ and $\lambda_h > 0$ imply $c^h > 0$ in the solution to (10) whenever $c > 0$. This also implies $u_\lambda^{*\prime}(0) = +\infty$ holds. Of course, if $\lambda_h = 0$, then $c^h = 0$ in the solution to (10).

The Pareto optimal growth model is then given by the classic discounted Ramsey model:

$$\sup \sum_{t=1}^{\infty} \delta^{t-1} u_\lambda^*(c_t) \tag{11}$$

subject to $c_t + k_t \leq f(k_{t-1})$, $t = 1, 2, \ldots$, c_t, $k_{t-1} \geq 0$, $k_0 \leq k$.

This problem has a unique solution under our basic assumptions. The neoclassical optimal growth model's properties obtain for this Pareto optimal growth model: the optimal aggregate consumption and capital sequences are monotonic and converge to the modified golden-rule consumption, c^*, and capital, k^*. Notice that the steady state capital stock and aggregate consumption levels are independent of the welfare weights. However, given c^*, the steady state allocations to the various households do depend on those weights by way of the solution to (10) with $c = c^*$. Different weights will distribute the steady state aggregate consumption differently. Consumption is equally distributed in the steady state if and only if the welfare weights are equal

with $\lambda h = 1/H$. Along dynamic equilibrium paths aggregate consumption growth also implies each household's consumption grows provided that agent's welfare weight is positive.

The preservation of the capital monotonicity property in this Pareto optimal growth problem suggests that the orthodox vision applies to its equilibrium counterpart. It turns out that with many agents the form of the equivalence principle is more subtle than with a single, representative, agent. The essential issue is the same problem that arises with the classical welfare theorems in finite dimensional commodity spaces – a Pareto optimum may only be a competitive equilibrium with transfer payments. Once this problem is handled, the basic equivalence principles carry over to the many agent case provided all households discount future utility at the same rate. The orthodox vision prevails.

The orthodox vision's realization in the Pareto optimal growth problem with equal discount factors does not extend to a model with heterogeneous agents and distinct discount factors. In this case, the household with the largest discount factor is the most patient one. The modified golden-rule capital stock, k^*, is still well-defined. However, Le Van and Vailakis (2003) prove the Pareto optimal capital sequence initiated at k^* converges to it in the long-run — but it is not a constant sequence: if the economy starts with the stocks k^*, then it is optimal for the planner to deviate from those stocks and only return to them asymptotically. The resulting optimal capital sequence cannot be monotonic, although the authors show it can be eventually monotonic. In part, this reflects the fact that the households enjoy timevarying consumption along their optimal path. The aggregate consumption levels change over time, but the first household emerges as the dominant consumer in the limit. The heterogeneous agent extension of the neoclassical representative agent theory does not exhibit the orthodox vision.

The Ramsey Equilibrium Model

The Ramsey equilibrium developed in Becker (1980) and reviewed in Becker (2006) interprets Ramsey's original long-run steady state conjecture with heterogeneous agents in a modern fashion. The basic model is developed for the case of agents with time additively separable utility functions with fixed discount factors. Each agent has a different discount factor, so one household is more patient than all the others. The technology is specified by a one-sector model with a single all-purpose consumption–capital good as before.

The general complete market competitive one-sector model treats budget constraints as restricting the present value of an agent's consumption to be smaller than or equal to the agent's initial wealth defined as the capitalized wage income plus the present value of that person's initial capital. This allows us to interpret the choice of a consumption stream as if the agent were allowed to borrow and lend at market-determined present value prices subject to repaying all loans. Markets are complete – any intertemporal trade satisfying the present value budget constraint is admissible at the individual level. The Ramsey equilibrium model changes the budget constraint from a single one reckoned as a present value to a sequence, one for each period. Agents are forbidden to borrow against their future labour income, so they cannot capitalize the future wage stream into a present value. Markets are incomplete. It becomes crucial to track the evolution of each person's capital stock. This is unnecessary in the complete market models when all values entering the budget constraint are present values.

The incomplete market structure shows itself in an individual's budget constraint. At each time, a household's available income is derived from rental returns on its capital stocks, and its wage rate (all labour is alike and inelastically supplied). Expenditure at each time is for consumption goods and for capital goods to be carried over to the next period in order to earn rental income. The borrowing constraint takes the form of a non-negativity constraint on the capital stock holdings in each time period. The formal constraint is analogous to (8) with superscripts attached to individual consumption and capital holdings.

The heterogeneous discount factor, incomplete market economy, differs in another important

respect: the operation of a borrowing constraint in the individual household problems also breaks the possibility of an equilibrium allocation arising as the economy's optimal allocation. The welfare maximization approach favoured in the complete market theory is inapplicable.

The Ramsey model has a unique stationary equilibrium in which only the most patient household has capital. That agent also enjoys a labour income. All other households consume their wages and own no capital. The model's dynamics have some distinctive features when compared with the capital and consumption monotonicity characteristic of the representative agent neoclassical model. The main results for the Ramsey equilibrium model appear in a series of papers beginning with Becker and Foias (1987). The survey article by Becker (2006) reviews those results as well as others in detail. Here, it is enough to note that the Ramsey equilibrium aggregate capital starting from an arbitrary distribution of initial capital stocks *eventually* has the capital monotonicity property in the case where the production function's elasticity of substitution is greater than or equal to 1, a condition satisfied by the Cobb–Douglas production function. In this case, the orthodox vision of capital eventually holds. If that elasticity of substitution condition fails, then Becker and Foias showed it was possible for a two-period equilibrium cycle to exist; the orthodox vision necessarily fails.

Behavioural Economics and Quasi-Geometric Discounting

The discounted Ramsey model where the planner discounts future utilities at a constant rate is the fundamental dynamic model in macrodynamics and economic growth theory. The time consistency of the optimal plan, based on the stationarity of the planner's utility function (even in the general recursive case) has been questioned by behavioural economics researchers on the basis of experiments and empirical evidence. For example, Ainslie (1991, p. 334) states that a majority of adults report they would rather have $50 immediately rather than $100 in two years, but almost no one chooses $50 in four years instead of receiving $100 in six years. If these individuals have stationary preferences, the mere passage of four years calendar time should not change the ranking of $50 in year four to $100 in year six if $50 was preferred in the present to $100 in two years. Thus, Ainslie concludes these individuals are time-inconsistent in their intertemporal preference ranking. Ainslie, as well as many others (notably Laibson, 1997; also see the survey by Frederick et al. (2002), for detailed summaries of the evidence and related references based on works by psychologists and economists) argue a different discounting function that describes real human behaviour better than the constant discounting model. The quasi-geometric discounting model developed below illustrates the simplest form of an alternative discounting function that these researchers argue better describes real human intertemporal choices. The quasi-geometric discounting function is an important example of the *hyperbolic discounting functions* appearing in behavioural discussions of time preference. The time preference reversals reported by Ainslie can be thought of as a criticism of standard discounted utility models in much the same way as the Allais paradox in risky choice experiments provides evidence against the expected utility model.

The standard constant discounting model's discounting function is $D(t) = \delta^{t-1}$, where $0 < \delta < 1$ is the discount factor and $t \geq 1$. The function D is also called the *exponential discount function*. The *quasi-geometric discounting* model posits a discounting function of the form $d(t) = \beta\delta^{t-1}$, where $\beta > 0$ is a parameter. The case $\beta = 1$ corresponds to the exponential discount function. If $\beta < 1$, there is short-run impatience – the decision maker is willing to save in the future, just not in the present. If $\beta > 1$, then there is short-run patience – the decision maker is more willing to consume in the future rather than the present. It is known from the fundamental paper by Strotz (1955) that, if a dynamic optimizing planner's discount factor does not have an exponential form, then the resulting optimal solution found from maximizing utility discounted to the present date will be time inconsistent. Thus, a planner solving the problem of maximizing the quasi-geometric utility function:

$$U(\mathbf{c}) = u(c_1) + \beta\left[\delta u(c_2) + \delta^2 u(c_3) + \cdots\right] \quad (12)$$

subject to (2) will exhibit time inconsistency. The solution $\{\overline{c}_t, \overline{k}_{t-1}\}_{t=1}^{\infty}$ so found will change if the planner is able to re-optimize at time 2. That new solution $\left\{\overline{c}^{\#}, \overline{k}^{\#}{}_{t-1}\right\}_{t=2}^{\infty}$ will have the property that $\overline{k}_2 \neq \overline{k}^{\#}{}_2$ when $\overline{k}_1 = \overline{k}_1{}^{\#}$ expresses the initial condition for the second period's optimization problem. Put differently, unless the planner can credibly commit to implementing the solution found in the first period, the planner will make another choice of optimal plans once period 2 is attained than the one originally found at time 1. The time inconsistent solution found in period 1 is really not an optmium as the planner would not implement it when called on to do so in the absence of a credible commitment to that plan.

Phelps and Pollak (1968) proposed a different way to arrive at a solution to the problem of maximizing (12) subject to (2). Their approach recognizes the planner must correctly anticipate future actions. The choice of c_t at some future date t alters the planner's capital stock and impacts the choices of consumption levels for all times past t. These impacts must be somehow considered by the planner in the present when the optimal plan is determined.

Phelps and Pollak imagined the planner as really infinitely many planners, each a *generation* that lives, saves and consumes over just one period. The discount factor, δ, measures impatience; the parameter β reflects the degree to which the current generation values future generations' utility relative to their own utility. Perfect altruism corresponds to the case $\beta = 1$ whereas imperfect altruism arises whenever $\beta < 1$. Later writers, following Laibson (1997), interpreted the generations as different *selves*, one for each time period. In either interpretation, the planner acts as if there are really infinitely many selves in the infinite-horizon Ramsey-styled optimization problem. Phelps and Pollak go on to argue the Ramsey optimal growth problem should be considered as a game with the many selves as the players. A Nash equilibrium of this game constitutes a solution to the planner's problem in the sense that no self (or generation) can improve its payoff given the actions taken by future selves (generations). Modern game theory research published after Phelps and Pollak's article suggests that such a game might have many equilibrium points. One possibility is the Markov perfect equilibrium concept. A Markov perfect equilibrium is time consistent. At time t, no histories of past choices or measurement of the capital stock are assumed to matter for outcomes beyond the current value of the aggregate capital stock that is presented to the self active at that moment. Other equilibrium notions can be formulated to reflect the game's history as play unfolds over time. Trigger strategies provide one way to do this. Of course, a fundamental equilibrium existence question arises for Markov perfect equilibrium as well as those equilibrium concepts derived from the selves adopting trigger strategies.

A Markov perfect equilibrium is represented by a time independent capital policy function, $g(k)$, that the current self expects to govern all future selves' saving and capital accumulation decisions. In this way, the aggregate capital stock is expected to evolve according to the dynamical system $k_t = g(k_{t-1})$ with $k_0 = k$, the capital stock endowment available at time 0. Note that this function depends only on the currently available capital stock. To solve the planner's quasi-geometric utility optimization problem is to find such a policy function. Recall that a policy function of this type characterized the solution to the canonical version of the discounted Ramsey model and reflected the underlying time consistency property of the planner's stationary utility function. It is also a Markov perfect equilibrium in the quasi-geometric case where $\beta = 1$ and $u(c) = \ln c$ with $f(k) = k^{\alpha}$. Of course, a major technical problem is to show a Markov perfect equilibrium exists in models where $\beta \neq 1$. For the log utility, Cobb–Douglas production model, a Markov perfect equilibrium has been constructed in the quasi-geometric case with $\beta < 1$ by Krusell et al. (2002). They showed that there is a Markov perfect equilibrium with policy function:

$$k' = g(k) = \frac{\beta\delta\alpha}{1-\alpha\delta(1-\beta)}k^{\alpha}, \qquad (13)$$

a functional form that agrees with the canonical example's capital policy function when $\beta = 1$. Iteration of this capital policy function (13) from the given initial capital stock produces a monotonic aggregate capital sequence. The qualitative properties of this particular Markov perfect equilibrium in this parameterized quasi-geometric model is the same as the qualitative properties of the canonical discounted Ramsey model, even though the two models' quantitative properties differ. For example, the two models have different steady states. The similarity was noted in Barro's (1999) continuous time model; he dubbed this similarity an observational equivalence result as the two models could not be distinguished empirically on the basis of their qualitative features alone.

Efficient Programs

Programs which are optimal for the discounted Ramsey model as well as its more general recursive utility formulations have an important efficiency property: there is no other feasible consumption sequence that provides more consumption in at least one period and as much in any other when compared with the optimal consumption path. This efficiency property can be studied in capital accumulation models in its own right as a minimal requirement for any reasonable objective function. Considered on its own, the efficiency criterion does not do much to single out a specific course of action for the planner. However, it can be used to eliminate some candidate optima without further reference to a specific welfare function. Moreover, examining efficient programs of consumption and capital accumulation can be undertaken in models with infinitely lived agents as well as models with finitely lived, overlapping generations where the economy evolves over an infinite horizon.

The interest in intertemporal efficiency stems from Malinvaud's (1953) seminal paper. He presented the first extension of Koopmans' activity analysis of efficient allocations in a static production world to an open-ended economy with a recursive technological structure, such as the aggregative one-sector model. He was also the first to recognize that the analog of Koopmans' profit conditions for characterizing an efficient program had to be supplemented in an infinite framework. This new terminal condition, the *transversality condition* (seen in the above discussion of the optimal growth model) was shown to be sufficient for an efficient program satisfying the profit conditions for an appropriate set of shadow prices.

Efficient programs are discussed below for the aggregative one-sector model. A sequence $\{c_t\}$ satisfying (2) for some capital stock sequence is *inefficient* if there is an alternative consumption program $\{c'_t\}$ satisfying (2) for some capital stock sequence that offers at least as much consumption in every period and more consumption in at least one period. A sequence $\{c_t\}$satisfying (2) for some capital stock sequence is *efficient* if it is not inefficient. The *efficiency criterion* ranks programs as either efficient or inefficient. The planner's objective is to select an efficient program. The efficiency criterion presumes that consumption may never be satiated in any period. An infinite number of efficient programs exists in the discounted Ramsey model – for a fixed, finite, time period T, define a feasible program by consuming nothing for periods $t = 0, 1, \ldots, T-1$, and letting the capital stock accumulate according to the difference equation $k_t = f(k_{t-1})$, with $k_0 = k$. At time T, consume the resulting $f(k_{T-1})$ and set $k_T = 0$. For each time after T, consume zero and accumulate no capital. Such a path is efficient. Since T is arbitrary, there are infinitely many efficient paths.

Efficient programs providing consumption in every period also exist. One important example is the path found by first solving for the combination of consumption and capital stock which maximizes stationary (or, sustainable) consumption. This program solves the problem $\max\{f(x) - x : x \in [0, b]\}$. The solution, denoted k^g, satisfies $f'(k^g) = 1$ and called the *golden-rule capital stock*; the corresponding *golden-rule consumption*, c^g, is defined by the relation $c^g = f(k^g) - k^g$. The interpretation is that if the economy's initial capital stock happens to equal

the golden-rule stock, then it is efficient for the planner to choose this stock for all time and maintain the largest possible stationary consumption.

The golden-rule pair (c^g, k^g) has an important relationship to the problem of characterizing efficient programs. The specific result is called the Phelps theorem (see Phelps, 1966, p. 59). It is a sufficient condition for an attainable path to be inefficient. A $\{c_t, k_t\}$ satisfying (2) also satisfies the *Phelps condition* if there is an $\varepsilon > 0$ and a natural number $T(\varepsilon)$ such that for all $t \geq T(\varepsilon)$, $k_t \geq k^g + \varepsilon$. The Phelps condition is equivalent to $\liminf_{t \to \infty} k_t > k^g$. Phelps' theorem states that a feasible program satisfying the Phelps condition is inefficient. In particular, the path of pure accumulation found by iterating $k_t = f(k_{t-1})$ for all t with $k_0 = k$ is inefficient as this program converges to the maximum sustainable capital stock. Any feasible program for which the capital stocks converge to a stock larger than the golden-rule stock is also inefficient. Note that such a program would have the own rate of return, $f'(k_{t-1}) - 1 < 0$ for all t sufficiently large. In particular, this would imply $\Pi_{t=1}^{T} f'(k_{t-1}) \to 0$as $T \to \infty$. It turns out that this is a general property of inefficient programs, as shown by Cass (1972). Intuitively, these inefficient programs have shadow interest rates, $r_t = f'(k_{t-1}) - 1$ that are negative (no market mechanism is identified in this discussion, so the interpretation of $f'(k_{t-1}) - 1$is provisionally made as a shadow price). It is reasonable then to presume that programs with positive shadow interest rates for all time are efficient. The precise criterion that is necessary and sufficient to characterize inefficient programs was identified by Cass (1972). He proved his result with additional curvature assumptions on the production function (which restrict the rate of change of capital's marginal product as capital accumulates, or decumulates) as well as assumed $f'(0) < \infty$. His theorem states that a feasible path is inefficient if and only if:

$$\sum_{t=1}^{\infty} \left[\prod_{s=1}^{t} f'(k_{s-1}) \right] < \infty.$$

Notice that if a path satisfies this Cass condition, then $\Pi_{s=1}^{t} f'(k_{s-1}) \to 0$ as $t \to \infty$, which is the Phelps sufficient criterion for inefficiency. Cass interpreted his condition as saying that the term $\Pi_{s=1}^{t} f'(k_{s-1})$goes to zero 'sufficiently fast'. The term $\Pi_{s=1}^{t} f'(k_{s-1})$ represents the shadow future value of a marginal unit of capital in period 0. The Cass criterion's necessity then asserts that for an inefficient program, the future value of a marginal unit of capital at time 0 is bounded from above. This implies that the terms of trade from present to future never become very favorable (Cass, 1972, p. 207). General forms of the Cass criterion for one-sector models are discussed in the survey by Becker and Majumdar (1989) as well as additional applications to overlapping generations models and interpretations of these conditions for decentralized planning mechanisms. The survey by Tirole (1990) focuses on the connection between the Cass criterion for inefficient programs and the potential for the shadow prices associated with efficient programs to exhibit a type of bubble whereby the shadow market price of a unit of capital differs from its present discounted value of future shadow rental returns.

Controversies and Critiques

Neoclassical capital theory has long been controversial. The famous *Cambridge Controversies* about whether or not the one-sector neoclassical model's properties were either sensible, or could be generalized, produced a substantial literature. See Birner (2002) for a thorough review of both sides' positions. Earlier references include Harcourt (1972), Bliss (1975), and Burmeister (1980). A few key points are noted here.

The debates centred on whether or not there really is something called aggregate capital, whether or not it could be measured independently of the establishment of an equilibrium interest rate, and whether or not an increase in the steady state interest rate necessarily reduced steady state capital.

Bliss (1975) argued that aggregating capital was not more difficult than aggregating any other collection of commodities. It was enough to place a partial order on a vector of capital goods

defining one vector of capital goods to be at least as much as another vector. Standard utility function existence theorems would imply the existence of a continuous, real-valued, order preserving functional representation that could be interpreted as an aggregate capital good. Burmeister (1980) gave conditions under which a generalized steady state regularity condition applied to a many capital goods model permitted theorists to construct an aggregate capital stock and aggregate production function with the desired neoclassical properties (at least across steady states). It should also be noted that there are models where there is a natural measure of an aggregate capital stock in physical terms. For example, the capital stock in renewable resource theories such as ones arising in fishery models measures the fish population as a *biomass*: the mass of living organisms present in a population at a particular point of time. Biomass can be measured as either a weight or as so many calories. Its measurement does not depend on any prices or other quantities that might be established only in an equilibrium. Of course, this is a special situation.

One practical way of arriving at a measure of aggregate capital is to compute its capital value. This can be done by multiplying the prices of the various underlying capital goods times their respective quantities. Presumably, these prices represent these capital goods' discounted future returns (for example, monetary or cash flows). Capitalization of future payments requires an interest rate (or a term structure of interest rates in case the rate of interest varies over time). It follows that capital value cannot be computed independently of the determination of prices. Critics of neoclassical theory stressed this issue. Modern equilibrium models establish the determination of capital goods prices and interest rates in an equilibrium configuration, for both the short and the long runs (this is one task solved by equivalence principles in many capital goods models, when those results are available).

The comparative steady state result for the one-sector neoclassical model is that the steady state capital stock, $k(\delta)$, viewed as a function of the discount (long-run interest) factor δ^{-1}, has the property $dk/d\delta > 0$. The famous reswitching controversy attacks the generality of this result. In multi-sectoral models (even with aggregate capital) the choice of steady state production techniques can give rise to a particular capital–labour ratio arising from two different long-run interest rates.

The Cambridge controversies highlight the special features of the one-sector neoclassical theory. Those arguments concentrated on comparing steady states and either ignored or downplayed the role for transitions from one steady state to another in response to an exogenous change in an economy's deep taste or technology parameters. The debate also largely ignored the accumulation programs that flowed from the planner's decision when starting with initial capital other than the steady state level. The more dynamic view of modern capital theorists emphasizes the full dynamic possibilities open to the planner.

The orthodox vision applied to an aggregative economy portrays saving and consumption activities undertaken within the private sector as promoting a path of accumulation tending towards a steady state. When the economy's capital stock is initially smaller than its stationary level there is growth, and the rate of return on capital falls over time. This portrait of capital accumulation is consistent with the dynamics of the one-sector Ramsey optimal growth – perfect foresight equilibrium model provided there is a representative household whose preferences are taken as the planner's objective.

Bliss (1975) criticized the orthodox vision for models with many distinct capital goods as a single rate of interest could not be defined, and therefore the idea that growth accompanied a declining rate of interest made no sense. Subsequent research has shown that, even in aggregate capital Ramsey optimal growth models with a well-defined interest rate, the economy might not follow the orthodox vision provided there were at least two sectors producing a consumption good distinct from the capital good. The problem was that optimal cycles or even chaotic trajectories could emerge with a sufficiently impatient planner (see Boldrin and Woodford, 1990). Heterogeneous discount factor models also turn out to

differ fundamentally from the representative agent theory, even in the classical one-sector case. The orthodox vision will only apply to *some* economies when there are heterogeneous discount factors.

The Cambridge controversy focuses on the difficulties of aggregating different types of capital and consumption goods. There are also difficulties inherent in interpreting results obtained for representative agent economies. The failure of the orthodox vision noted above is one such example. There is another, perhaps more fundamental, criticism of representative agent-based capital theories. The conditions under which the many different individuals populating a model economy's preferences might be aggregated so that the economic theorist can study the model as if there is a single, stand-in, representative agent are so restrictive as to make conclusions drawn from single agent models flawed on logical grounds alone. See Hartley (1997) for a detailed discussion of the representative agent controversy.

The idea of a representative agent economy such as the Ramsey model is that the aggregate activity in the economy generated by many different consuming and producing actors can be understood as the activity of a single entity, the representative agent, which acts exactly like each of the consuming and producing actors. By studying the microeconomic behaviour of those individuals we can also find the behaviour of the representative agent, and vice versa. However, the argument is made that, even if the microfoundations of each agent are well understood, it does not follow that their aggregate behaviour is explained by the representative agent that behaves exactly like them. Micro-behaviour need not translate into macro-behaviour of the same type. For example, the representative agent Ramsey model's capital monotonicity property holds up in the welfare optimum version of the many agent theory when agents have the same discount factors, but different one-period utility functions and possibly different initial capital stocks. The planner whose preferences are represented by the welfare function (11) does not give rise to the exact same behaviour as that of each of the individual agents' preferences underlying it – individual consumption sequences differ from the aggregate, although they behave qualitatively the same (for example, they are monotonic). This distinction is even more pronounced in case agents also have different discount factors – the impatient agents' consumption tends to zero while the most patient one's consumption remains positive for all time. The aggregate consumption evolves over time in a very different manner from that of individual consumption streams.

Capital Theory with Many Sectors and Capital Goods

Controversies surrounding the neoclassical capital theory of the one-sector model are partly attenuated by studying models with many sectors and types of capital goods. This general form of the theory emphasizes a disaggregated viewpoint, although it also applies to aggregative models. It should also be noted that specifying a multisector model need not be the same as formulating a many capital good model. There are two-sector models with aggregative capital and single-sector models with joint production of many distinct capital and consumption goods.

Pricing and the Portfolio Equilibrium Condition

The major conceptual difference between the one-sector and multisector perfect foresight equilibrium models lies in the form taken by the no-arbitrage condition. This is readily seen in the two-sector model. Suppose there are two sectors consisting of a consumption goods sector and a capital goods sector. The capital and consumption goods are aggregate commodities, as in the one-sector model, but are conceived as distinct goods in the two-sector framework. Suppose that i_{t+1} is the one-period interest rate measured in units of a numeraire commodity, r_{t+1} is the rental rate on a unit of capital measured in the numeraire's units, and q_{t+1} is the unit purchase price for a unit of capital as measured in the numeraire's units. Suppose that the purchase of a unit of capital at time t entitles its owner to receive the rental flows from the next period on as long as

the unit remains in service. Assume further that capital does not depreciate. One requirement for a perfect foresight equilibrium is that there are no one-period reversed arbitrage opportunities. Let an equilibrium path obtain with the prices $\{i_{t+1}, r_{t+1}, q_{t+1}\}$. Suppose the household decision maker acquires another unit of capital at time t. This costs the household q_t units of the numeraire. The opportunity cost of this action in the numeraire's units is $i_{t+1}q_t$, the interest charge that could have been earned otherwise. To reverse this capital acquisition at time $t+1$ the household will sell that unit of capital for q_{t+1} units of the numeraire. This gives the capital gain (loss) equal to $q_{t+1} - q_t$. The household also gets to keep the one-period rental, r_{t+1}. This one-period reversed arbitrage is unprofitable if the marginal revenue equals the marginal cost reckoned in units of the numeraire. That is,

$$i_{t+1}q_t = r_{t+1} + q_{t+1} - q_t. \tag{14}$$

This equation reflects the absence of arbitrage opportunities in a perfect foresight competitive equilibrium. This *perfect foresight equation* is also called the *portfolio equilibrium condition* because it expresses the absence of arbitrage opportunities in the manner in which the agent's wealth is held. Rearranging this equation yields

$$i_{t+1} = \frac{r_{t+1}}{q_t} + \frac{q_{t+1} - q_t}{q_t}, \tag{15}$$

which says that the one-period interest rate, i_{t+1}, equals the capital good's *own rate of return*, r_{t+1}/q_t, plus the *capital gain yield*, $(q_{t+1} - q_t)/q_t$.

Note that $q_t = 1$ holds in the one-sector model. This is the price of the consumption good in units of the numeraire commodity (chosen to be current consumption) since the capital and consumption goods are identical. Hence, there is no capital gain yield in that case and

$$i_{t+1} = r_{t+1}. \tag{16}$$

The interest rate equals the rental rate for capital goods. Thus, even if there is a single capital good, the portfolio equilibrium condition differs when the one-sector and two-sector models are compared.

Next, consider an aggregate model with an exhaustible resource. Suppose there are neither extraction nor storage costs. The aggregate capital stock at the end of time period t that is available for consumption at time $t+1$ is denoted by k_t and is interpreted as the amount of the resource remaining at the end of time t.

Consumption at time t, ct, represents a withdrawal from the stock k_{t-1}. Then the materials balance condition is $c_t + k_t = k_{t-1}$. The initial size of the resource stock is k. There is no rental return in this model; the resource owner's returns are entirely capital gain yields. The perfect foresight equation takes the form

$$i_{t+1} = \frac{q_{t+1} - q_t}{q_t}. \tag{17}$$

If the rate of interest is a constant: $i_{t+1} = r > 0$, then (17) is a linear difference equation with solution $q_{t+1} = (1+r)^t q_0$, where q_0 is the resource's initial price. This implies *Hotelling's r-per cent rule* (Hotelling, 1931) holds in a perfect foresight equilibrium – the equilibrium (current) price of the resource, q_t, increases over time at rate of interest, r.

In models with several distinct capital goods the portfolio equilibrium condition applies to each capital good separately. If there are m capital goods, then the portfolio equilibrium condition takes the form:

$$i_{t+1} = \frac{r^j_{t+1}}{q^j_t} + \frac{q^j_{t+1} - q^j_t}{q^j_t}, \text{for} j = 1, 2, \ldots, m. \tag{18}$$

Here, the superscript j labels capital good j. With many capital goods households have a variety of options for holding their wealth. The rates of return on any portfolio of capital stocks must be equalized or there will be a one-period reversed arbitrage opportunity. Hence, eq. (18) is the equilibrium condition expressing the absence of such arbitrage opportunities.

The major pricing differences between the one-sector and multisector models concern the form of the portfolio equilibrium condition. It is possible to develop equivalence principles for multisector models along the lines of the one-sector theory by making appropriate adjustments in the pricing of capital goods to reflect their multiplicity in the budget constraints and production sector while also recognizing the portfolio equilibrium form of the no-arbitrage conditions in the PFCE and FCE settings.

Establishing the formal equivalence between optimal accumulation models and their equilibrium counterparts in many capital good models requires the equilibrium economy to impose a transversality condition on itself, just as in the one-sector case. The general question is how is the initial price determined so that the equilibrium price profile satisfies the conditions for achievement of a Ramsey-styled central planning solution. This is the crux of the *Hahn problem*. The modern perfect foresight interpretation is that this problem is solved whenever a transversality condition obtains as necessary for an equilibrium. This requires the household sector to be forward looking over the infinite horizon, and markets to operate on all dates and for all commodities. Some writers on capital theory take a critical view of these conditions and argue that markets cannot be relied on to set the correct initial prices, and so the resulting equilibrium path is inefficient. On the other hand, a comparison of idealized markets with idealized planning, as embodied in the equivalence principles, suggests that at the most theoretical level the Hahn problem is resolved when rational, forward-looking agents conduct their economic activities in a complete market setting over an infinite horizon.

Final Comments

The constraints of the neoclassical one-sector model can be used to substitute for consumption in the felicity function by noting $u(c_t) = u(f(k_{t-1}) - k_t)$, where $c_t \geq 0$ if and only if $f(k_{t-1}) - k_t \geq 0$. The current period's payoff depends only on the stocks of capital at the beginning and end of the period. This observation results in a reformulation of the one-sector model focused on the capital stock sequences. Let $u(0) = 0$ to simplify the exposition. Let $D = \{(x, y) \in \mathbb{R}_+ \times \mathbb{R}_+ : f(x) - y \geq 0\}$. Note that $(0,0) \in D$. The felicity function $v(x, y) \equiv u(f(k_{t-1}) - k_t)$ has domain D and $v(0, 0) = 0$. The properties of u and f imply that v is increasing in its first argument and decreasing in its second argument. The concavity of u and f also imply that v is a concave function defined on the convex set D. The planner continues to discount future utility by the factor δ, $0 < \delta < 1$. This alternative representation of the neoclassical model, called the *reduced form model*, gives rise to an optimal growth problem with the planner choosing the sequence $\{k_t\}_{t=0}^{\infty}$ to achieve

$$\sup_{\{k_t\}_{t=0}^{\infty}} \sum_{t=1}^{\infty} \delta^{t-1} v(k_{t-1}, k_t); (k_{t-1}, k_t) \in D \text{ for each } t, \text{ and } 0 \leq k_0 \leq k. \tag{19}$$

This form of the one-sector model is just one realization of the general reduced form model. A complete exposition of this general structure's properties is found in McKenzie's surveys. The reduced form model can accommodate many varieties of capital theoretic problems including multisector and multi-capital good models, von Neumann's model of economic growth, exhaustible and renewable resource models, as well as individual firm investment theory when there are costs of adjusting the firm's capital stocks. The capital stocks of the one-sector model are replaced by a vector of capital stocks where each component represents a particular capital good; the set D is then contained in a multi-dimensional Euclidean space. Schefold (1997) is a recent treatment of multisector models derived from Sraffa's (1960) perspective on capital accumulation models that also revisits the reswitching controversy in a dynamic equilibrium setting. Also see Burmeister (1980) for a critical exposition of Sraffa's contribution. Burgstaller (1995) reviews models from the Sraffa tradition as well as neoclassical models in continuous time in order to find their common ground and connections to earlier capital theories.

The full scope of capital theoretic problems in deterministic, continuous time can be found in Weitzman (2003). The monograph by Becker and Boyd (1997) addresses the analogous problems in discrete time. Conrad and Clark (1987) covers natural resource models from a dynamic perspective. Stokey et al. (1989) provide an excellent introduction to stochastic dynamic models along with development of the discrete time theory using dynamic programming techniques. Chang (2004) presents basic continuous time stochastic calculus and optimal control theory with economic applications including the classical tree-rotation problem.

See Also

- ▶ Capital Theory (Paradoxes)
- ▶ Dynamic Programming
- ▶ Intertemporal Equilibrium and Efficiency
- ▶ Neoclassical Growth Theory (New Perspectives)
- ▶ Present Value
- ▶ Ramsey Model

Bibliography

Ainslie, G. 1991. Derivation of 'rational' economic behavior from hyperbolic discount curves. *American Economic Review* 81: 334–340.

Barro, R.J. 1999. Ramsey meets Laibson in the neoclassical growth model. *Quarterly Journal of Economics* 114: 1125–1152.

Beals, R., and T.C. Koopmans. 1969. Maximizing stationary utility in a constant technology. *SIAM Journal of Applied Mathematics* 17: 1001–1015.

Becker, R.A. 1980. On the long–run steady state in a simple dynamic model of equilibrium with heterogeneous households. *Quarterly Journal of Economics* 95: 375–382.

Becker, R.A. 2006. Equilibrium dynamics with many agents. In *Handbook of Optimal Growth Theory, Volume I: Discrete Time Theory*, ed. C. Le Van, R.A. Dana, T. Mitra, and K. Nishimura. New York: Springer-Verlag.

Becker, R.A., and J.H. Boyd. 1997. *Capital Theory, Equilibrium Analysis, and Recursive Utility*. Malden, MA: Blackwell Publishers.

Becker, R.A., and C. Foias. 1987. A characterization of Ramsey equilibrium. *Journal of Economic Theory* 41: 173–184.

Becker, R.A., and M. Majumdar. 1989. Optimality and decentralization in infinite horizon economies. In *Joan Robinson and Modern Economic Theory*, ed. G.R. Feiwel. London: Macmillan.

Birner, J. 2002. *The Cambridge Controversies in Capital Theory: A Study in the Logic of Theory Development*. London: Routledge.

Bliss, C.J. 1975. *Capital Theory and the Distribution of Income*. Amsterdam: North-Holland/America Elsevier.

Boldrin, M., and M. Woodford. 1990. Equilibrium in models displaying fluctuations and chaos: a survey. *Journal of Monetary Economics* 25: 189–222.

Brock, W.A., and L.J. Mirman. 1972. Optimal growth and uncertainty: the discounted case. *Journal of Economic Theory* 4: 479–513.

Burgstaller, A. 1995. *Property and Prices: Toward a Unified Theory of Value*. Cambridge: Cambridge University Press.

Burmeister, E. 1980. *Capital Theory and Dynamics*. Cambridge: Cambridge University Press.

Cass, D. 1972. On capital overaccumulation in the aggregative model of economic growth: a complete characterization. *Journal of Economic Theory* 4: 200–223.

Chang, F.R. 2004. *Stochastic Optimization in Continuous Time*. Cambridge: Cambridge University Press.

Conrad, J.M., and C.W. Clark. 1987. *Natural Resource Economics: Notes and Problems*. Cambridge: Cambridge University Press.

Dixit, A.K. 1976. *The Theory of Equilibrium Growth*. Oxford: Oxford University Press.

Dorfman, R., P.A. Samuelson, and R.M. Solow. 1958. *Linear Programs and Economic Analysis*. New York: McGraw-Hill.

Epstein, L.G. 1983. Stationary cardinal utility and optimal growth under uncertainty. *Journal of Economic Theory* 31: 133–152.

Epstein, L.G., and J.A. Hynes. 1983. The rate of time preference and dynamic economic analysis. *Journal of Political Economy* 91: 611–635.

Fisher, I. 1907. *The Rate of Interest*. New York: Macmillan.

Frederick, S., G. Loewenstein, and T. O'Donoghue. 2002. Time discounting and time preference: a critical review. *Journal of Economic Literature* 40: 351–401.

Goetzmann, W.N., and K.G. Rouwenhorst. 2005. *The Origins of Value: The Financial Innovations That Created Modern Capital Markets*. Oxford: Oxford University Press.

Harcourt, G.C. 1972. *Some cambridge controversies in the theory of capital*. Cambridge: Cambridge University Press.

Hartley, J.E. 1997. *The representative agent in macroeconomics*. London: Routledge.

Hotelling, H. 1931. The economics of exhaustible resources. *Journal of Political Economy* 39: 137–175.

Koopmans, T.C. 1958. *Three essays on the state of economic science*. New York: McGraw–Hill.

Krusell, P., B. Kuruscu, and A.A. Smith. 2002. Equilibrium welfare and government policy with quasigeometric discounting. *Journal of Economic Theory* 105: 42–72.

Laibson, D. 1997. Golden eggs and hyperbolic discounting. *Quarterly Journal of Economics* 112: 443–477.
Le Van, C., and Y. Vailakis. 2003. Existence of a competitive equilibrium in a one–sector growth model with heterogeneous agents and irreversible investment. *Economic Theory* 22: 743–771.
Lucas, R.E., and N.L. Stokey. 1984. Optimal growth with many consumers. *Journal of Economic Theory* 32: 139–171.
Malinvaud, E. 1953. Capital accumulation and efficient allocation of resources. *Econometrica* 21: 233–268.
McKenzie, L.W. 1986. Optimal economic growth, turnpike theorems and comparative dynamics. In *Handbook of mathematical economics*, ed. K. Arrow and M.D. Intriligator, Vol. 3. Amsterdam: North-Holland.
McKenzie, L.W. 1987. Turnpike theory. In *The New Palgrave: A dictionary of economics*, ed. J. Eatwell, M. Milgate, and P. Newman, Vol. 4. London: Macmillan.
Mirman, L.J., and I. Zilcha. 1975. Optimal growth under uncertainty. *Journal of Economic Theory* 11: 329–339.
Phelps, E.S. 1966. *Golden rules of economic growth*. New York: Norton.
Phelps, E.S., and R. Pollak. 1968. On second-best national saving and gameequilibrium growth. *Review of Economic Studies* 35: 185–199.
Radner, R. 1961. Paths of economic growth that are optimal with regard only to final states. *Review of Economic Studies* 28: 98–104.
Rae, J. 1834. *Statements of some new principles on the subject of political economy*. Reprinted 1964. New York: Augustus M. Kelley.
Ramsey, F.P. 1928. A mathematical theory of saving. *Economic Journal* 38: 543–559.
Schefold, B. 1997. *Normal prices, technical change and accumulation*. London: Macmillan.
Solow, R.M. 1956. A contribution to the theory of economic growth. *Quarterly Journal of Economics* 70: 65–94.
Sraffa, P. 1960. *Production of commodities by means of commodities*. Cambridge: Cambridge University Press.
Stokey, N.L. and Lucas, R.E., with Prescott, E. 1989. *recursive methods in economic dynamics*. Cambridge, MA: Harvard University Press.
Strotz, R.H. 1955. Myopia and inconsistency in dynamic utility maximization. *Review of Economic Studies* 11: 165–180.
Tirole, J. 1990. Intertemporal efficiency, intergenerational transfers, and asset pricing: an introduction. In *Essays in honor of Edmond Malinvaud, vol I: Microeconomics*, ed. P. Champsaur et al. Cambridge, MA: MIT Press.
von Neumann, J. 1937. Über ein Ökonomisches Gleichungssystem une eine Verallgemeinerung des Brouwerschen Fixpunksatzes. *Ergebnisse eines mathematischen Kolloquiums* 8, 73–83. *Review of Economic Studies* 13(1945): 1–9.
Weitzman, M.L. 2003. *Income, wealth, and the maximum principle*. Cambridge, MA: Harvard University Press.

Capital Theory (Paradoxes)

Luigi L. Pasinetti and Roberto Scazzieri

Abstract

Capital theory has led economists to discover relationships that look 'paradoxical' or counter-intuitive, as they run counter widely accepted 'parables'. The transformation of microeconomic diminishing returns relations into a macro-social law induced the mistaken belief of an inverse, monotonic relation between the interest rate (and profit rate, taken as the 'price of capital') and the quantity of capital per head. Subsequent work alerted economists to the difficulty of finding aggregate measures of heterogeneous capital goods, and to the possibility that a falling rate of interest (and of profit) may be associated with a decrease not an increase) of the quantity of capital per head.

Keywords

Aggregation (of capital); Aggregation (production); Aghion, P.; Beccaria, C.; Bliss, C. J.; Böhm-Bawerk, E. von; Bruno, M.; Burmeister, E.; Cantillon, R.; Capital accumulation; Capital as a collection of physical assets; Capital as a fund; Capital intensity; Capital measurement; Capital theory; Capital theory: parables; Capital theory: paradoxes; Chain index method of measurement; Champernowne, D. G.; Clark, J. B.; Classical capital theory; Cobb–Douglas functions; Cohen curiosum; Cohen, R.; Construction period and utilization period, in Hicks's theory of capital; Denison, E.; Diminishing returns; Discontinuities in input use; Distribution of income; Duration parameters in production models; Embodied technical change; Financial and technical conceptions of capital; Fisher, F.; Garegnani, P.; Gordon, R.; Harcourt, G. C.; Heterogeneity of capital goods; Hicks, J. R.; Hidden reswitching; Howitt, P.; Hulton, C.;

Increasing returns; Intertemporal equilibrium; Jevons, W. S.; Kaldor, N.; Law of variable proportions; Levhari, D.; Loanable funds; Locke, J.; Longfield, M.; Marginal productivity theory; Morishima, M.; Multifactor productivity; Neoclassical production function; Net capital stocks; Non-monotonic relations in capital theory; Pasinetti, L.; Petty, W.; Physiocracy; Production matrix; Production techniques; Rate of interest; Rate of profit; Reswitching of technique; Reverse capital deepening; Ricardo, D.; Robinson, J. V.; Rosser, J. B.; Roundabout methods of production; Rymes, T.; Samuelson, P. A.; Say, J.-B.; Schefold, B.; Senior, N.W.; Sheshinski, E.; Smith, A.; Solow, R.; Sraffa, P.; Starrett, D.; Steady state; Stiglitz, J.; Structural change; Structural economic dynamics; Surrogate production function; Sylos Labini, P.; Temporary equilibrium; Thünen, J. H. von; Variable proportions, law of; Vertically integrated sectors; Vertically integrated labour coefficients; Vertically integrated units of productive capacity; Weitzman, M.; Wicksell, K.

JEL Classifications
E22

The idea that capital theory might lead economists to discover forms of 'paradoxical' behaviour emerged in the economic literature of the 1960s largely as an outcome of developments in the field of production theory (linear production models leading to enquiries into discrete and discontinuous relations). What happened in capital theory is in fact a special instance of a more general phenomenon. Economists sometimes tend to examine a large domain of economic phenomena by adapting theoretical concepts that had originally been devised for a much narrower range of special issues. The discoveries of 'paradoxical' relations derive from the fact that their process of generalization often turns out to be ill-conceived and misleading, if not entirely unwarranted.

For a long time, in capital theory it had been taken for granted that there is a unique, unambiguous profitability ranking of production techniques in terms of capital intensity, along the scale of variation of the rate of interest. The discovery that this is not necessarily true has induced many economists to speak of 'paradoxes' in the theory of capital. But the roots of apparently paradoxical behaviour are to be found, not in the economic phenomena themselves, but in the economists' tendency to rely on too simple 'parables' of economic behaviour.

Traditional beliefs about capital are deeply rooted in the history of economic analysis, and may be traced back to pre-classical literature. As will be shown in the next section, a long post-classical tradition was then developed on that basis. The length of ancestry might explain the survival of conventional beliefs.

The Emergence of the Conventional View

The notion of 'capital' was associated for a long time with investible wealth and its income generating power, and was largely independent of detailed consideration of the function of invested wealth in the production process. The earliest development of capital theory took place within the accounting framework of a pre-industrial economy (William Petty, John Locke, Richard Cantillon). Within this perspective, capital was often associated with purely financial transactions (lending and borrowing) and the relationship between capital and rate of interest came quite naturally to be conceived as the relationship between loanable funds and their price (see Cannan 1929, pp. 122–53). The origin of the belief in an inverse monotonic relation between the demand for capital and the rate of interest may be traced back to this phase of the literature. The distinction between capital as a fund of purchasing power and capital as a 'sum of values' embodied in physical assets remained in the background (see Hicks 1977, p. 152), but was bound, in time, to generate tension 'between the physical and financial conceptions of capital' (Cohen and Harcourt 2005, p. xli).

The association of capital with the process of production did not come to the fore until quite late, in spite of certain isolated anticipations. (John Hicks 1973, p. 12, even quotes Boccaccio's

Decameron on the issue.) The description of capital as a stock of means of production became common with the Physiocrats and the classical economists. In this period, Cesare Beccaria (1804, ms 1771–72) presented what Jean-Baptiste Say considered to be the first analysis of 'the true functions of productive capitals' (Say 1817, p. xliii). Soon after him, Adam Smith (1776) built upon the distinction between 'productive' capital and 'unproductive' consumption his theory of structural dynamics and economic growth. Finally, David Ricardo gave a definite shape to classical capital theory by examining the relationship between capital accumulation and diminishing returns and by considering in which way different proportions of capital in different industries might influence the relative exchange values of the corresponding commodities (Ricardo 1817, ch. 1, sections 4 and 5).

Classical capital theory is characterized by lack of interest in the purely financial dimension of investment. As a result, the relation between capital accumulation and the rate of interest recedes into the background and is substituted by the relation between real capital accumulation and the rate of profit. In this way, the foundations of capital theory shifted from the exchange to the production sphere, and the demand-and-supply mechanism was confined to the process by which the rate of interest is maintained equal to the rate of profit in the long run. However, a number of economists (starting with Johann Heinrich von Thünen, Mountifort Longfield and Nassau William Senior) continued to be interested in the income-generating function of capital at the level of the individual investor, and tried to combine this approach with the emphasis on the productive function of capital that had emerged in the classical literature. The marginal productivity theory of capital and interest was developed as an answer to this conceptual problem. The essential features of that theory may be clearly seen in Thünen, who suggested a relationship between the rate of interest (i) and the rate of profit (r) quite different from the one found in Ricardo. The reason for this is that Ricardo had taken r to be fixed for the individual entrepreneur, so that equality between i and r was brought about by adjustment between the supply and demand for loans in the financial markets. Thünen suggested a different adjustment mechanism by taking r to be variable for the individual entrepreneur, so that the attainment of the long-run equality between the rate of profit and the rate of interest came to depend on the change in the physical productivity of capital as much as on adjustment in the financial markets (see Thünen 1857).

This view is founded upon a thorough transformation of the Ricardian theory of diminishing returns and provided the logical starting point for the later marginalist theory of diminishing returns from aggregate capital. The analytical and historical process leading to this outcome is a rather complex one, and it is best understood by distinguishing two separate stages. In the first stage, the law of diminishing returns, which Ricardo considered to hold for the economy as a whole in the long run, was applied to the short-run behaviour of the individual entrepreneur. As result, the change in input proportions within any given productive unit is associated with the change in the physical productivity of capital. Here the variation of the capital stock is unlikely to influence the system of prices, so that the decrease (or increase) in the return from the last 'increment of capital' could be unambiguously associated with an increase (or decrease) in the physical capital stock. The second stage consisted in extending the above result to the variations in the aggregate quantity of capital available in the economic system as a whole.

The process which we have described made it possible to transform the classical conception of diminishing returns from a macro-social law into a microeconomic relation derived from the law of variable proportions. This new type of diminishing returns was then extended to the 'macro-social' sphere once again. As a result, it became possible to think that the rate of interest and the rate of profit (tending to be equal to each other) are associated with the physical marginal productivity of aggregate capital: an increase in the relative quantity of capital with respect to the other inputs would be associated with lower marginal productivity of capital and thus with a lower equilibrium rate of interest and rate of profit. This

inverse monotonic relation between the rate of interest (and the rate of profit) and the quantity of capital per head eventually became an established proposition of capital theory. The relevance of this relation can be seen from the attempts by William Stanley Jevons (1871), Eugen von Böhm-Bawerk (1889) and John Bates Clark (1899) to found on the theory of the marginal productivity of factors the explanation of the distribution of the social product among factors of production under competitive conditions.

Further light on the conceptual roots of the marginalist view of capital is shed by the contributions of Jevons and Böhm-Bawerk. In their theories, profit is considered as the remuneration due to the capitalist as a result of the higher productiveness of 'indirect' or 'roundabout' processes of production than of processes carried out by 'direct' labour only. The generalization of the marginal principles which they carried out is thus associated with the description of the production process as an essentially 'financial' phenomenon in which final output, like interest in financial transactions, could be considered as 'some continuous function of the time elapsing between the expenditure of the labour and the enjoyment of the result' (Jevons 1879, p. 266). The subsequent discovery of 'anomalies' in the field of capital accumulation was possible when economists started to question this extension of capital theory from the financial to the productive sphere, and when the technical structure of production was examined on its own grounds independently of the 'financial' aspect which might be considered to be characteristic of 'the typical business man's viewpoint' (Hicks 1973, p. 12).

Anticipations of Debate

It has just been shown that microeconomic diminishing returns provided the foundations for a theory of the diminishing marginal productivity of social capital, which was extended from the microeconomic sphere by way of logical analogy.

The pitfalls of this approach did not take long to emerge, as economic analysis came to grips with the full complexity of the production process. Knut Wicksell, discovered that, in the case of an economic system using heterogeneous capital goods, it might be impossible to describe diminishing returns from aggregate capital. The reason for this is that a variation in the capital stock might be associated with a change in the price system that would make it impossible to compare the quantities of capital before and after the change (see Wicksell 1901–6, pp. 147 ff. and 180). Wicksell also recognized that this difficulty is characteristic of capital because 'labour and land are measured each in terms of its own *technical* unit … capital, on the other hand, … is reckoned, in common parlance, as a sum of *exchange value'* (1901–6, p. 149).

The special difficulty associated with heterogeneous capital goods is in fact an outcome of a particular procedure by which the fundamental theorems concerning capital and interest had been formulated with reference to the idealized setting of an isolated producer, and then extended by analogy to the case of the 'social economy'. The drawbacks of this methodology were perspicaciously noted by Nicholas Kaldor in the late 1930s, when he complained that capital theory had been developed starting with 'a … specialised set-up, with the picture of Robinson Crusoe engaged in net-making' rather than with the 'general case' of 'a society where *all* resources are produced and the services of all resources co-operate in producing further resources' (Kaldor 1937, p. 228.) Kaldor also noted that, had the analysis started with the 'general case', 'a great deal of the controversies concerning the theory of capital might not have arisen' (Kaldor 1937, p. 228).

It is remarkable that so many 'paradoxical' results of modern capital theory were subsequently discovered precisely as an outcome of the procedure here described by Kaldor.

The stage of modern controversy was set by the consideration of two distinct problems: (*a*) the measurement of 'aggregate capital' in models with heterogeneous capital goods; and (*b*) the discovery that production techniques that had been excluded at lower levels of the rate of profit might 'come back' as the rate of profit is increased (this phenomenon is known as *reswitching of technique*).

Joan Robinson started the discussion by calling attention to the difficulties inherent in any physical measure of aggregate capital (Robinson 1953–4). She also pointed out the 'curiosum' that the degree of mechanization associated with a higher wage rate and a lower rate of profit might be lower than the degree of mechanization associated with a lower wage rate and a higher rate of profit. (She attributed this 'curiosum' to Miss Ruth Cohen, but later on she attributed it to her reading of Sraffa's Introduction to Ricardo's *Principles*.)

Immediately afterwards, David Champernowne discovered that, in general, we must admit 'the possibility of two stationary states each using the same items of equipment and labour force yet being shown as using different quantities of capital, merely on account of having different rates of interest and of food-wages' (Champernowne 1953–4, p. 119). Champernowne also admitted that the inverse monotonic relation between the rate of profit and the quantity of capital per head (as well as the inverse monotonic relation between the rate of profit and capital per unit of output) might not be generally true: 'it is logically possible that over certain ranges of the rate of interest, a fall in interest rates and rise in food-wages will be accompanied by a *fall* in output per head and a *fall* in the quantity of capital per head' (Champernowne 1953–4, p. 118). Champernowne's explanation of what appeared to be perverse behaviour from the point of view of traditional theory was that changes in the interest rate can be associated with changes in the cost of capital equipment even if the physical capital stock is unchanged. As a result, perverse behaviour was attributed to pure 'financial' variations and a physical measure of capital was still thought to be possible. This Champernowne tried to obtain by introducing a chain index method for measuring capital (Champernowne 1953–4, p. 125). A few years later, Joan Robinson again took up the same issue in her *Accumulation of Capital* (1956, pp. 109–10). The reason she gave for the 'Ruth Cohen curiosum' is quite different from the one proposed by Champernowne. She explicitly recognized that 'financial' factors such as a higher wage rate and a lower rate of interest would have 'real' consequences by influencing the actual choice of technique. (In the 'perverse' case a lower rate of interest would be associated with the choice of the less mechanized technique.)

When a few years later Michio Morishima attempted a multi-sectoral generalization of Joan Robinson's simple model he confirmed the possibility of a positive relationship between the rate of interest and the degree of mechanization of a technique (Morishima 1964, p. 126). Finally John Hicks came up with the same problem when examining 'the response of technique to price changes' in the framework of a simple economy consisting of a consumption good 'industry' and a net investment good 'industry', and in which the same capital good is used in both industries (see Hicks 1965, pp. 148–56).

But, in spite of all these anticipations, it must be admitted that the issue of technical reswitching was not given an important place in economic theory before the publication of Piero Sraffa's *Production of Commodities by Means of Commodities* (1960). It is with Sraffa's work that the phenomenon took a prominent place. Sraffa was able to show that heterogeneity of capital goods and of 'capital structures' (different proportions between labour and intermediate inputs in the various processes of production) would normally give rise, with the variation of the rate of profit and of the unit wage, 'to complicated patterns of price-movement with several ups and down' (Sraffa 1960, p. 37). This phenomenon would in turn bring about changes in the 'quantity of capital' that are not generally related to the rate of profit in a monotonic way. Reswitching of technique and reverse capital deepening are thus derived from a general property of production models with heterogeneous capital goods. (See reswitching of technique and reverse capital deepening.)

Neoclassical Parables and the Capital Controversy

Following the publication of Sraffa's book, a lively debate on capital theory suddenly flared up in the 1960s, and the way it did is itself an interesting event.

It has already been pointed out that, when propositions derived from individual behaviour are applied to the more complex case of the 'social economy', the extension is admittedly possible on condition that the social economy has a number of special features making it identical, from the analytical point of view, to the case of the isolated individual. To test these features, the social economy is often described in terms of a 'parable' in which those particular conditions are satisfied. This 'parable', though unrealistic, is taken to be useful, from an heuristic or a persuasive point of view.

In this vein Paul Samuelson attempted to construct a 'surrogate production function' by analogy with microeconomic behaviour (Samuelson 1962). His work can be considered as the first explicit attempt to get rid of the complexities of an economic system with heterogeneous capital goods by constructing a model in which that system is described in terms of an 'aggregate parable' with physically homogeneous capital. After introducing the assumption that 'the same proportion of inputs is used in the consumption-goods and [capital-] goods industries' (Samuelson 1962, pp. 196–7), Samuelson was able to prove that 'the Surrogate (Homogeneous) Capital ... gives exactly the same result as does the shifting collection of diverse capital goods in our more realistic model' (1962, p. 201). In particular, 'the relations among w, r, and Q/L that prevail for [the] quasi-realistic complete system of heterogeneous capital goods' could 'be shown to have the same formal properties as does the parable system' (1962, p. 203). This result was taken to be a justification for using the surrogate production function 'as a useful summarizing device' (1962, p. 203). In fact, Pierangelo Garegnani, who was present at a discussion of a draft of Samuelson's paper, did point out that Samuelson's result is crucially dependent on the assumption of equal proportions of inputs (see Garegnani 1970). Samuelson acknowledged Garegnani's criticism in a footnote to his paper and admitted that it would be a 'false conjecture' to think that the 'extreme assumption of equi-proportional inputs in the consumption and machine trades could be lightened and still leave one with many of the surrogate propositions' (Samuelson 1962, p. 202n). But Samuelson and various other economists continued to look for conditions that would ensure a monotonic relation between the rate of profit and the choice of technique even in presence of a nonlinear relation between w and r.

The outcome appeared a few years later. David Levhari, a Ph.D. student of Samuelson's, in his dissertation and then in a paper for the *Quarterly Journal of Economics,* claimed he had proved that reswitching of the whole production matrix would be impossible if this matrix is of the 'irreducible' or 'indecomposable' type (Levhari 1965). This property – Levhari claimed – would exclude reswitching and thus make it possible to extend the use of a 'surrogate production function' to the nonlinear case with production technologies for basic commodities.

However, Levhari's theorem was disproved by Luigi Pasinetti in a paper at the Rome First World Congress of the Econometric Society in 1965. Pasinetti's final draft of his paper was published in the November 1966 issue of the *Quarterly Journal of Economics* (Pasinetti 1966) together with papers written in the meantime by David Levhari and Paul Samuelson (1966), Paul Samuelson (1966), Michio Morishima (1966), Michael Bruno et al. (1966) and Pierangelo Garegnani (1966). This set of papers was called by the journal editor 'Paradoxes in Capital Theory: A Symposium', thereby originating the term. Paul Samuelson concluded the discussion with a 'Summing up' in which he admitted that 'the simple tale told by Jevons, Böhm-Bawerk, Wicksell, and other neoclassical writers', according to which a falling rate of interest is unambiguously associated with the choice of more capital-intensive techniques, 'cannot be universally valid' (Samuelson 1966, p. 568).

The various contributions to this discussion showed that reswitching might occur both with 'decomposable' and 'indecomposable' technology matrices. This result was proved in different ways by Pasinetti (1965, 1966), Morishima (1966), Bruno et al. (1966) and Garegnani (1966). Samuelson stated in his summing up that 'reswitching is a logical possibility in any technology, indecomposable or decomposable' (1966,

p. 582). He then called attention to the associated phenomenon of reverse capital deepening and concluded that 'there often turns out to be no unambiguous way of characterizing different processes as more "capital-intensive", more "mechanized", more "roundabout"' (1966, p. 582).

Although the logical possibility of reswitching was admitted by all participants in the discussion, Bruno, Burmeister and Sheshinski raised doubts as to its empirical relevance: 'there is an open empirical question as to whether or not reswitching is likely to be observed in an actual economy for reasonable changes in the interest rate' (Bruno et al. 1966, p. 545n). The same doubt was expressed in Samuelson's summing up (Samuelson 1966, p. 582). Bruno, Burmeister and Sheshinski also mentioned a theorem, which they attributed to Martin Weitzman and Robert Solow, according to which reswitching of technique may be excluded, in a model with heterogeneous capital goods, provided at least one capital good is produced by 'a smooth neoclassical production function', if 'labour and each good are inputs in one or more of the goods produced neoclassically' (Bruno et al. 1966, p. 546). This theorem is based on the idea that 'setting the various marginal productivity conditions and supposing that at two different rates of interest the same set of input–output coefficients holds, the proof follows by contradiction' (Bruno et al. 1966, p. 546).

It is worth noting that Weitzman–Solow's theorem is simply a consequence of the idea that, in the case of a commodity produced by a neoclassical production function, each set of input–output coefficients ought to be associated in equilibrium with a one-to-one correspondence between marginal productivity ratios and input price ratios. No ratio between marginal productivities would be associated with more than one set of input prices, and this is taken to exclude the possibility that the same technique be chosen at alternative rates of interest, and thus at different price systems. The Weitzman–Solow theorem is at the origin of a line of arguments that has been followed up by a number of other authors, such as David Starrett (1969) and Joseph Stiglitz (1973). These authors have pursued the idea that 'enough' substitutability, by ensuring the smoothness of the production function, is sufficient to exclude reswitching of technique. However, non-reswitching theorems of this type involve that, for each technique of production, the capital stock may be measured either in physical terms or at given prices. For in a model with heterogeneous capital goods, if we allow prices to vary when the rate of interest or the unit wage are changed, there is no reason why the same physical set of input–output coefficients might not be associated with different price systems: even in the case of a continuously differentiable production function, the marginal product of 'social' capital cannot be a purely real magnitude independent of prices. Once it is admitted that 'in general marginal products are in terms of net value at constant prices, and hence may well depend upon what those prices happen to be' (Bliss 1975, p. 195), it is natural to allow for different marginal productivities of the same capital stock at different price systems. It would thus appear that reswitching of technique does not carry with it any logical contradiction even in the case of a smoothly differentiable production function.

But Pasinetti also pointed out that the concept of neoclassical substitutability is itself a very restrictive concept indeed, as it requires the possibility of infinitesimal variations of each input at a time. In fact, Pasinetti noted that it is possible to have a continuous variation of techniques (that is, continuous substitutability) along the w–r relation and yet wide discontinuities in the variation of many inputs between one technique and another, thus making reswitching a quite normal phenomenon (see Pasinetti 1969). Moreover, and even more significantly, a non-monotonic relation between the rate of profit and capital per man may well be obtained even in the absence of reswitching (Pasinetti 1966; Bruno et al. 1966). This last possibility calls attention to the phenomenon that lies at the root of the various 'paradoxes' in the theory of capital: the fact that, unless special assumptions are made, a change in the rate of profit and in the unit wage at given technical coefficients is associated with a change of relative prices.

This debate continued for a few years in the late 1960s and early 1970s, with a series of journal

articles (see for example Robinson and Naqvi 1967) and books (see for example Harcourt 1972). In particular, John Hicks presented a 'Neo-Austrian' model in *Capital and Time* (1973), concluding that reswitching of technique can be excluded only in the special case in which all the techniques have the same 'duration parameters', which means the same 'construction period' and 'utilization period' (1973, pp. 41–4).

In the end, numerous details were added. Yet the basic essential results remained those that had come out of Sraffa's book and of the symposium on 'Paradoxes in Capital Theory'. It is instructive to see that, in a recent exchange of views that has appeared in the *Journal of Economic Perspectives* (2003, Spring and Winter issues), Franklin Fisher (2003), Geoff Harcourt in Cohen and Harcourt (2003) and Luigi Pasinetti (2003), when asked to succinctly summarize the issues at stake, have essentially restated their original positions.

Aftermath and Ways Ahead

The discovery of paradoxes in capital theory has had a number of important repercussions, mostly beyond its original context. For it stimulated a large amount of analytical and empirical research on some of the issues that had been discussed in the controversy, without pressing the attention towards the fundamentals, as had been the case with the original debates. In many instances, the recent developments have been motivated by the need to face the problem of measuring the stock of capital goods in economic systems subject to advances of technical knowledge and structural change, or some of the associated issues in the theory of economic dynamics. In this section we shall refer to some of these developments without pretending to give a complete picture, but with the purpose of identifying the main lines of inquiry.

A first area of research has been the analysis of the necessary conditions for the empirical measurement of aggregate capital. Franklin Fisher elaborated a research line he had himself started in an earlier contribution (Fisher 1969) and called attention to the fact that the aggregation of outputs, as well as that of productive factors, 'requires separability in each firm's production function' (Fisher 1987, p. 55). He also noted that, under constant returns, the two highly restrictive assumptions of no specialization and generalized capital augmentation are necessary, whereas, in most cases of non-constant returns, aggregation would not be allowed even when assuming the same production function for all firms (Fisher 1987, p. 55). Robert Gordon proposed to measure collections of heterogeneous capital goods, under condition of embodied technical change, by considering the associated 'net revenue at a given set of prices (w) of variable inputs' (Gordon 1993, p. 106; see also Gordon 1990). Edward Denison did find Gordon's proposal objectionable and proposed instead to 'equate' new capital goods with the old ones by 'what their relative costs *would* be if both *were* produced at a common date' (Denison 1993, pp. 89–90). An interesting link between this literature and the capital controversy debate has been suggested by Charles Hulton, who has called attention to the advantages of a 'recursive description of the production possibility set', in which the assumption of capital as an original input is dropped, and 'capital and labour are assumed to produce gross output *and* capital which is one period older' (Hulten 1992, p. S15). Hulten's formulation highlights the central role of knowledge advances embodied in new capital goods and suggests the relevance, for distinct purposes, of gross outputs and net outputs 'as indicators of capacity and economic welfare' (Hulten 1992, p. S11). Alexandra Cas and Thomas Rymes have specifically addressed the issue of whether 'knowledge of the constant-price aggregate stock of capital would, *for the comparison of economies,* permit one to "predict" certain variables' (Cas and Rymes 1991, p. 7; emphasis added). In particular, they investigated capital measurement issues brought about by embodied technical change, and proposed a set of 'new measures' aimed at taking the fact into account that 'the net capital stocks of each industry and at the aggregate are themselves being produced with increased efficiency when the capital goods industries are experiencing advances in technical knowledge' (Cas and Rymes 1991, p. 67). The same authors

relate their measures of changing capital stocks under conditions of structural change to 'Pasinetti's concepts of vertically integrated sectors and productivity aggregated by end use' (Cas and Rymes 1991, pp. 90–1). This point of view highlights the common ground behind recent attempts to measure stocks of heterogeneous capital goods in terms of an aggregate concept of productive capacity, be it Pasinetti's 'unit of vertically integrated productive capacity' (Pasinetti 1973, 1981), Cas and Rymes' 'new measures of multifactor productivity' (Cas and Rymes 1991), or Hulten's 'accounting for capacity' (Hulten 1992). In all these cases, the producibility of capital goods is emphasized, as is the close connection between advances of technical knowledge and the reshuffling of inter-industry relationships (particularly those affecting intermediate goods). Philippe Aghion and Peter Howitt have commented on recent discussions about capital measurement for an economy subject to advances of knowledge by recalling Joan Robinson's view that the real issue is not so much about the measurement of capital as rather about the *meaning* one wishes to assign to any given collection of capital goods (Aghion and Howitt 1998, p. 435).

Another line of investigation has concerned the attempt to assess the empirical (or computational) relevance of capital paradoxes, as distinct from their theoretical possibility. In this connection, Stefano Zambelli has used computer simulations in order to investigate the 'realism' of capital paradoxes in artificial economies (Zambelli 2004). This author has found a significantly higher likelihood that the capital–labour ratio be *positively* related to the rate of profit, contrary to the conventional belief of a negative relationship between these two variables. This result is consistent with the empirical investigation carried out by Zonghie Han and Bertram Schefold (2006). These authors have compared pairs of techniques from the OECD input–output database, and have found that 'observed cases of reswitching and reverse capital deepening are more than flukes' (Han and Schefold 2006, p. 22), even if we are far from observing what has been called an 'avalanche of switchpoints' (Schefold 1997, pp. 278–80).

A third line of research has carried the discussion of capital paradoxes into the field of dynamic economic theory. The literature relevant in this connection is itself quite differentiated. For example, Frank Hahn (1966) called attention to his earlier discovery of zones of instability in economies with heterogeneous capital goods, and pointed out that reswitching should be considered as one amongst the multiple causes of instability in capital markets (Hahn 1982). It is interesting that this line of argument, while maintaining that reswitching is a special case of a larger class of phenomena, at the same time and rather surprisingly also makes reswitching to be *more general* than was the case with earlier treatments of the same phenomenon. For capital paradoxes are no longer mainly associated with an economy with heterogeneous capital goods and a uniform rate of profit, but are ' extended' to the case of multisectoral economies with many different capital goods and a *multiplicity* of rates of interest (and rates of profit). Luigi Pasinetti followed a different approach, and examined the analytical features of a dynamic economy in which market interactions are not explicitly examined (Pasinetti 1981). In this case, too, there are reasons to think that reswitching and reverse capital deepening would not represent exceptional cases, and would not be limited to the institutional framework of a perfectly competitive economy. Other authors have examined the relationship between capital paradoxes and dynamic stability, and have argued that reswitching of technique and reverse capital deepening are neither necessary nor sufficient conditions for the economic system to show *lack* of stability and irregular behaviour (Mandler 2005). It has also been emphasized that 'reswitching' adds an important element of instability, the importance of which depends on the process of adaptation, but also on the utility function' (Schefold 2005, p. 467).

More generally, the discovery of capital paradoxes has stimulated a deeper understanding of the features of continuity and discontinuity in the dynamics of economic systems. This line of research has its point of departure in a phenomenon detected by Luigi Pasinetti shortly after the climax of the controversy (Pasinetti 1969). In

Pasinetti's more recent words, 'the vicinity, even the infinitesimal vicinity, of any two techniques on the scale of variation of the rate of profits does not entail at all vicinity of such techniques ... discontinuities in input use.' (Pasinetti 2000, p. 409). John Barkley Rosser Jr. has picked up such suggestions and has investigated the discontinuities in order to identify the implications of capital paradoxes for the analysis of the optimal dynamic path followed by an economy characterized by 'an infinite, differentiable technology' (Rosser 1983, p. 182). This author acknowledges that it may sometimes be impossible to directly observe reswitching along optimal adjustment path (as maintained, for example in Burmeister and Hammond 1977), but he notes that this would only happen 'at the price of dynamic discontinuities', that is, on the condition that the economic system be able to 'jump over' the zone associated with intermediate techniques. The above result has been interpreted as showing that 'in a world of infinite and smooth technologies, reswitching is to be "observed" by observing discontinuities in optimal dynamic paths' (Rosser 1983, p. 183; see also Rosser 2000, pp. 213–20). This point of view emphasizes the analytical importance of capital paradoxes as characteristic instances of the discontinuities that may be generated by the nonlinearity of certain structural relationships. In this way, the propositions discovered during the capital controversies of the mid-20th century are found to be consilient with much later developments in the economic analysis of nonlinear dynamic systems.

Synthesis

The source of most of the difficulties that have emerged in capital theory may be traced back to the fact that 'capital' may be conceived in two fundamentally different ways: (a) as a 'free' fund of resources, which can be switched from one use to another, without any significant difficulty: this is what may be called the 'financial' conception of capital; (b) as a set of productive factors that are embodied in the production process as it is carried out in a particular productive establishment: this is what may be called the 'technical' conception of capital.

The idea that there exists an inverse monotonic relation between the rate of interest and the demand for capital was born in the financial sphere. The parallel idea of an inverse monotonic relation between the rate of profit and the 'quantity of capital' employed in the production process is the outcome of a long intellectual process of extensions and generalizations reviewed earlier in this essay. But the recent debate on capital theory has conclusively proved that such extensions and generalizations are devoid of any foundation. It is logically impossible to make the 'financial' and the 'technical' conceptions of capital coincide, except under very restrictive conditions indeed. More precisely, there is no unambiguous way in which a decreasing rate of profit may be related to the choice of alternative techniques, in terms of monotonically increasing capital intensity, be this considered in terms of capital per unit of output or of capital per unit of labour.

These analytical results are hardly in dispute by now. But their ultimate significance and relevance for economic theory have been, and remain, controversial.

A group of economists have been so impressed by the new discoveries in capital theory, concerning the relations between rate of profit, capital per head, capital per output, and technical progress, as to become convinced that these discoveries are calling for a reconstruction of economic theory from its very foundations. It is stressed that the traditional beliefs are due to mistaken generalizations from the theory of short-run microeconomic behaviour, and it is argued that the economic theory ('marginal economic theory') that led to mistakes and inconsistencies should be abandoned. It is also pointed out that the obvious alternative is a resumption and development of the more comprehensive approach to value, distribution and growth of the classical economists (see Garegnani 1970, 2005, and, in a different context, Pasinetti 1981).

A second line of interpretation maintains that economic theorists should be prepared to give up the analytical tools of equilibrium analysis and

concentrate much more on the actual historical dynamics of economic systems. In this vein, reswitching of technique is acknowledged as a logical possibility but doubts are expressed on its importance in actual economic history (see Robinson 1975, pp. 38–9; Hicks 1979, p. 57).

A third line of interpretation is taken by more traditionally minded theoretical economists. It is argued that the discovery of 'anomalies' in the field of capital theory does point to an important deficiency in 'marginal' economic theory, which leads to the inevitable abandonment of the concept of 'aggregate capital'. However, it is also argued that there is a way of overcoming this deficiency without giving up the basic premises of traditional theory, and in particular without rejecting the application of the demand-and-supply framework to the study of production. This way induces to concentrating the analysis either on the study of 'short-run' ('temporary') equilibria, in which the physical stocks of capital are given, or on the equilibrium of an intertemporal economy, in which goods are described by taking their dates of delivery into account. In either case, the logical possibility (or 'existence') of an equilibrium price vector is studied without explicitly considering the movement of 'free' capital from one use to another. In this approach, the importance of 'capital paradoxes' is explicitly recognized, but the associated difficulties are transferred either to the field of stability analysis or to the theory of the long-period supply of saving as financial capital (see, respectively, Hahn 1982; Bliss 2005).

A fourth line of interpretation has been pursued by many empirically oriented economists. It is acknowledged that the notion of 'aggregate' technical capital is untenable in terms of theory, but it is also argued that the utilization of aggregate production functions may be justified on pragmatic terms, due to supposedly satisfactory econometric fit (see, for example, Fisher 1971; Fisher et al. 1977). This view however, is by no means widely accepted. It has in fact been vigorously challenged by Paolo Sylos Labini (1995), who has reviewed the estimates that have emerged from using the Cobb–Douglas production function and has shown that such a 'production function, when estimated econometrically, tends to yield, in general, poor results' (Felipe and Fisher 2003, p. 251; see also McCombie 1998; and Felipe and Adams 2005). In a recent evaluative essay on aggregation in production functions, Jesus Felipe and Franklin Fisher have sharply criticized the continued use of aggregate parables. In particular, they maintain that 'the revival of growth theory during the last two decades no doubt has produced important discussions, and seemingly interesting empirical results' but ' authors do not realize that they are using a tool whose lack of legitimacy was demonstrated decades ago' (Felipe and Fisher 2003, pp. 250–1). The same economists emphasize that 'the impossibility of testing empirically the aggregate production function' is 'substantially more serious than a mere anomaly', and that 'macroeconomists should pause before continuing to do applied work with no sound foundation and dedicate some time to studying other approaches to value, distribution, employment, growth, technical progress etc., in order to understand which questions can legitimately be posed to the empirical aggregate data' (Felipe and Fisher 2003, pp. 256–7). It is interesting that the theoretical and empirical researches that have taken up this challenge have devoted attention to the construction of a 'capacity measure' of the stock of technical capital that would allow comparisons across different states of technology without having recourse to the traditional 'parables' (see, for example, Pasinetti 1973, 1981; Cas and Rymes 1991; Hulten 1992).

Finally, let us note how the discovery of 'paradoxes' in capital theory has contributed to stimulating research into the dynamic properties of economic systems outside the world of steady state comparisons. In particular, some economists have attempted the theoretical investigation of regularities in the long-run dynamics of economic systems by suggesting a reformulation of the classical theory of structural change in a disaggregated framework (see Pasinetti 1981, 1993; Hagemann et al. 2003). Others have investigated the complex interaction of behavioural patterns along a dynamic trajectory, and have called attention to increasing returns and other

nonlinear phenomena in structurally adaptive economic systems (see Anderson et al. 1988; Arthur et al. 1997).

Whatever the view that is taken, the major victim of the debate has been the Böhm-Bawerk–Clark–Wicksell theory of capital that was so patiently constructed towards the end of the 19th century. This theory relied on a conception of 'aggregate capital' that was taken as measurable independently of the rate of profit and of income distribution. Such a conception of 'capital' has had to be jettisoned, which has stimulated reformulations of the pure theory of capital. There has been on the one hand a return to the Walrasian general equilibrium theory in its intertemporal formulation, and on the other hand a remarkable revival of classical political economy. The controversy had also a number of less striking but perhaps longer-term consequences. The consideration of paradoxes has alerted economists to the richness and complexity of economic relationships, and to the need to avoid a process of generalization from the consideration of special cases. In any case the debate seems to have compelled theoretical economists to be more rigorous about the nature and limits of their assumptions. In many important cases, it has also brought about a change in the main focus of their analysis.

All this leads one reasonably to expect as unlikely that the next generation of economists will leave the issue of capital theory at rest.

See Also

- ▶ Reswitching of Technique
- ▶ Reverse Capital Deepening

Bibliography

Aghion, P., and P. Howitt. 1998. *Endogenous growth theory*. Cambridge, MA: MIT Press.

Anderson, P., K.J. Arrow, and D. Pines, eds. 1988. *The economy as an evolving complex system*. Redwood City: Addison Wesley.

Arthur, W.B., S. Durlauf, and D. Lane, eds. 1997. *The economy as an evolving complex system II*. Redwood City: Addison-Wesley.

Baranzini, M., and R. Scazzieri. 1990. *The economic theory of structure and change*. Cambridge: Cambridge University Press.

Beccaria, C. 1804. Elementi di economia pubblica (ms 1771–72). In *Scrittori Classici Italiani di Economia Politica,* Vol. XVIII, 17–356 and Vol. XIX, 5–166, ed. P. Custodi. Milan: Destefanis.

Bliss, C.J. 1975. *Capital theory and the distribution of income.* Amsterdam/Oxford/North-Holland/New York: American Elsevier.

Bliss, C.J. 2005. Introduction. The theory of capital: A personal overview. In *Capital theory*, ed. J. Bliss, A.J. Cohen, and G.C. Harcourt, vol. 1. Cheltenham/Northampton: Edward Elgar.

Bruno, M., E. Burmeister, and E. Sheshinski. 1966. The nature and implications of the reswitching of techniques. *Quarterly Journal of Economics* 80: 526–553.

Burmeister, E., and P. Hammond. 1977. Maximin paths of heterogeneous capital accumulation and the instability of paradoxical steady states. *Econometrica* 45: 853–870.

Cannan, E. 1929. *A review of economic theory*. London: P.S. King and Son Ltd..

Cas, A., and T.K. Rymes. 1991. *On concepts and measures of multifactor productivity in Canada, 1961–1980*. Cambridge/New York: Cambridge University Press.

Champernowne, D. 1953. The production function and the theory of capital: A comment. *Review of Economic Studies* 21: 112–135.

Clark, J.B. 1899. *The distribution of wealth*. New York: Macmillan.

Cohen, A.J., and G.C. Harcourt. 2003. Retrospectives. Whatever happened to the Cambridge capital theory controversies? *Journal of Economic Perspectives* 17 (1): 199–214.

Cohen, A.J., and G.C. Harcourt. 2005. Introduction. Capital theory controversy: Scarcity, production, equilibrium and time. In *Capital theory*, ed. J. Bliss, A.J. Cohen, and G.C. Harcourt, vol. 1. Cheltenham/Northampton: Edward Elgar.

Denison, E.F. 1993. Robert J. Gordon's concept of capital. *Review of income and wealth* 39: 89–102.

Felipe, J., and F.G. Adams. 2005. A theory of production: The estimation of the Cobb–Douglas function: A retrospective view. *Eastern Economic Journal* 31: 427–445.

Felipe, J., and F.M. Fisher. 2003. Aggregation in production functions: What applied economists should know. *Metroeconomica* 54: 208–262.

Fisher, F.M. 1969. The existence of aggregate production functions. *Econometrica* 37: 553–577.

Fisher, F.M. 1971. Aggregate production functions and the explanation of wages: A simulation experiment. *Review of Economics and Statistics* 53: 305–325.

Fisher, F.M. 1987. Aggregation problem. In *The new Palgrave: A dictionary of economics*, ed. J. Eatwell, M. Milgate, and P. Newman, vol. 1. London: Macmillan.

Fisher, F.M. 2003. Cambridge capital controversies. *Journal of Economic Perspectives* 17 (4): 228–229.
Fisher, F.M., R. Solow, and J.M. Kearl. 1977. Aggregate production functions: Some CES experiments. *Review of Economic Studies* 44: 305–320.
Garegnani, P. 1966. Switching of techniques. *Quarterly Journal of Economics* 80: 554–587.
Garegnani, P. 1970. Heterogeneous capital, the production function and the theory of distribution. *Review of Economic Studies* 37: 407–436.
Garegnani, P. 2005. Capital and intertemporal equilibria: A reply to Mandler. *Metroeconomica* 56: 411–437.
Gordon, R.J. 1990. *The measurement of durable goods prices*. Chicago/London: University of Chicago Press.
Gordon, R.J. 1993. Reply: The concept of capital. *Review of Income and Wealth* 39: 103–110.
Hagemann, H., M. Landesmann, and R. Scazzieri, eds. 2003. *The economics of structural change*. Cheltenham/Northampton: Edward Elgar.
Hahn, F. 1966. Equilibrium dynamics with heterogeneous capital goods. *Quarterly Journal of Economics* 53: 633–646.
Hahn, F. 1982. The neo-Ricardians. *Cambridge Journal of Economics* 6: 353–374.
Han, Z., and B. Schefold. 2006. An empirical investigation of paradoxes: Reswitching and reverse capital deepening in capital theory. *Cambridge Journal of Economics* 30: 737–765.
Harcourt, G.C. 1972. *Some Cambridge controversies in the theory of capital*. Cambridge: Cambridge University Press.
Hicks, J. 1965. *Capital and growth*. Oxford: Clarendon Press.
Hicks, J. 1973. *Capital and time: A Neo-Austrian theory.* Oxford: Clarendon Press.
Hicks, J. 1977. Capital controversies: Ancient and modern. In *Economic perspectives: Further essays on money and growth*. Oxford: Clarendon Press.
Hicks, J. 1979. *Causality in economics*. Oxford: Basil Blackwell.
Hulten, C.R. 1992. Accounting for the wealth of nations: The net versus gross output controversies and its ramifications. *Scandinavian Journal of Economics* 94 (Supplement): 9–24.
Jevons, W.S. 1879. *The theory of political economy.* 2nd ed. London: Macmillan and Co (1st edn 1871).
Kaldor, N. 1937. Annual survey of economic theory: The recent controversy on the theory of capital. *Econometrica* 5: 201–233.
Levhari, D. 1965. A nonsubstitution theorem and switching of techniques. *Quarterly Journal of Economics* 79: 98–105.
Levhari, D., and P. Samuelson. 1966. The nonswitching theorem is false. *Quarterly Journal of Economics* 80: 518–519.
McCombie, J.S.L. 1998. Are there laws of production?: An assessment of the early criticisms of the Cobb–Douglas production function. *Review of Political Economy* 10: 141–173.
Mandler, M.A. 2005. Well-behaved production economies. *Metroeconomica* 56: 477–494.
Morishima, M. 1964. *Equilibrium,stability and growth: A multi-sectoral analysis*. Oxford: Clarendon Press.
Morishima, M. 1966. Refutation of the nonswitching theorem. *Quarterly Journal of Economics* 80: 520–525.
Pasinetti, L.L. 1965. Changes in the rate of profit and degree of mechanization: A controversial issue in capital theory. Unpublished paper presented at the First World Congress of the Econometric Society, Rome.
Pasinetti, L.L. 1966. Changes in the rate of profit and switches of techniques. *Quarterly Journal of Economics* 80: 503–517.
Pasinetti, L.L. 1969. Switches of technique and the 'rate of return' in capital theory. *Economic Journal* 79: 508–531.
Pasinetti, L.L. 1973. The notion of vertical integration in economic analysis. *Metroeconomica* 25: 1–29.
Pasinetti, L.L. 1981. *Structural change and economic growth: A theoretical essay on the dynamics of the wealth of nations*. Cambridge: Cambridge University Press.
Pasinetti, L.L. 1993. *Structural economic dynamics: A theory of the economic consequences of human learning*. Cambridge: Cambridge University Press.
Pasinetti, L.L. 2000. Critique of the neoclassical theory of growth and distribution. *Banca Nazionale del Lavoro Quarterly Review* 53: 383–431.
Pasinetti, L.L. 2003. Cambridge capital controversies. *Journal of Economic Perspectives* 17 (4): 227–228.
Ricardo, D. 1817. On the principles of olitical economy and taxation. Vol. I of *the works and correspondence of David Ricardo,* ed. P. Sraffa with the collaboration of M.H. Dobb, Cambridge: Cambridge University Press, 1951.
Robinson, J. 1953. The production function and the theory of capital. *Review of Economic Studies* 21 (2): 81–106.
Robinson, J. 1956. *The accumulation of capital*. London: Macmillan.
Robinson, J. 1975. The unimportance of reswitching. *Quarterly Journal of Economics* 89, 32–39. Reprinted in Robinson, J. 1979. *Collected economic papers,* Vol. 5. Oxford: Basil Blackwell.
Robinson, J., and K.A. Naqvi. 1967. The badly behaved production function. *Quarterly Journal of Economics* 81: 579–591.
Rosser, J.B. Jr. 1983. Reswitching as a cusp catastrophe. *Journal of Economic Theory* 31: 182–193.
Rosser, J.B. 2000. From Catastrophe to Chaos: A general theory of economic discontinuities. In *Mathematics, microeconomics, macroeconomics, and finance*, vol. 1, 2nd ed. Boston/Dordrecht/London: Kluwer Academic Publishers.
Samuelson, P.A. 1962. Parable and realism in capital theory: The surrogate production function. *Review of Economic Studies* 29: 193–206.
Samuelson, P.A. 1966. A summing up. *Quarterly Journal of Economics* 80: 568–583.

Say, J.B. 1817. *Traité d'économie politique*. 3rd ed. Paris: Déterville (1st edn 1803).
Schefold, B. 1997. *Normal prices, technical change and accumulation*. London: Palgrave Macmillan.
Schefold, B. 2005. Reswitching as a case of instability of intertemporal equilibria. *Metroeconomica* 56: 438–476.
Smith, A. 1776. An inquiry into the nature and causes of the wealth of nations. Vol.II of the glasgow edition of the works and correspondence of Adam Smith. Oxford: Clarendon Press. General editors R.H. Campbell and A.S. Skinner; textual editor W.B. Todd.
Sraffa, P. 1960. *Production of commodities by means of commodities*. Cambridge: Cambridge University Press.
Starrett, D. 1969. Switching and reswitching in a general production model. *Quarterly Journal of Economics* 83: 673–687.
Stiglitz, J. 1973. The badly behaved economy with the well-behaved production function. In *Models of economic growth*, ed. J.A. Mirrlees and N.H. Stern. London: Macmillan.
Sylos Labini, P. 1995. Why the interpretation of the Cobb–Douglas production function must be radically changed. *Structural Change and Economic Dynamics* 6: 485–504.
Thünen, J.-H. 1857. *Le salaire naturel et son rapport au taux de l'intérêt.* Trans. M. Wolkoff. Paris: Guillaumin et Cie. German original published in 1850.
von Böhm-Bawerk, E. 1889. *Positive Theorie des Kapitales* (*Kapital und Kapitalzins,* Zweite Ableitung). Trans. The Positive Theory of Capital. London: Macmillan.
Wicksell, K. 1901–6. *Lectures on political economy. Vol. 1, General theory.* London: Routledge, 1934.
Zambelli, S. 2004. The 40% neoclassical aggregate theory of production. *Cambridge Journal of Economics* 28: 99–120.

Capital Theory: Debates

Heinz D. Kurz

Capital theory is notorious for being perhaps the most controversial area in economics. This has been so ever since the very inception of systematic economic analysis. Much of the interest in the theory of capital lies in the fact that it holds the key to the explanation of profits. Since the notion of 'capital' is at the centre of an inquiry into the laws of production and distribution in a capitalist economy, controversies in the theory of capital are reflected in virtually all other parts of economic analysis.

We can distinguish between debates *within* different traditions of economic analysis and debates *between* them. In what follows our concern will be mainly with the latter. At the cost of severe simplification, the various traditions in the theory of capital and distribution may be divided into two principal groups, one rooted in the surplus approach of the classical economists from Adam Smith to Ricardo and the other in the demand and supply approach of the early marginalist economists. The so-called 'Cambridge controversies' (cf. Harcourt 1969), triggered off by a seminal paper by Joan Robinson (1953), consisted essentially in a confrontation of these two radically different traditions. The debate is still continuing. Currently, the discussion focuses on some of the neoclassical authors' claim that the classical theory, as it was reformulated by Sraffa (1960), is a 'special case' of modern general equilibrium theory. We shall come back to this questionable proposition towards the end of the entry.

The Surplus Approach

The general method underlying the classical economists' approach to the theory of capital and distribution was that of 'normal' or 'long-period' positions. These were conceived as centres around which the economy is assumed to gravitate, given the competitive tendency towards a uniform rate of profit. Because of the assumed gravitation of 'market values' to the 'normal' levels of the distributive and price variables, the former were given little attention only, being governed by temporary and accidental causes, a proper scientific analysis of which was considered neither necessary nor possible. Emphasis was on the persistent or non-temporary causes shaping the economy. Accordingly, the investigation of the permanent effects of changes in the dominant causes was carried out by means of comparisons between 'normal' positions of the economic system.

The development of a satisfactory theory to determine the general rate of profit was thus the

main concern of the classical economists. As regards the content of this theory, profits were explained in terms of the *surplus product* left after making allowance for the requirements of reproduction, which were conceived inclusive of the wages of labour (Ricardo 1817, vol. 1, p. 95). As Sraffa (1951, 1960) emphasized, the determination of the social surplus implied taking as data (i) the system of production in use, characterized, as it is, by the dominant technical conditions of production of the various commodities and the size and composition of the social product; and (ii) the ruling real wage rate(s). In accordance with the underlying 'normal' position the capital stock was assumed to be so adjusted to 'effectual demand' (Adam Smith) that a normal rate of utilization of its component parts would be realized and a uniform rate of return on its supply price obtained. Thus the classical authors separated the determination of profits and prices from that of quantities. The latter were considered as determined in another part of the theory i.e. the analysis of accumulation and economic and social development.

The rate of profit was defined by the ratio between social surplus and social capital, i.e. two aggregates of heterogeneous commodities. Thus the classical theory had to face the problem of value. Ricardo's ingenious device to solve this problem consisted in relating the exchange values of the commodities to the quantities of labour directly and indirectly necessary to produce them. This led to the first formulation of one of the key concepts in the theory of capital ever since – the inverse relationship between the real wage and the rate of profit (Ricardo, vol. 8, p. 194).

It was not until Marx that additional important steps in the development of the surplus approach were taken. In particular, in Marx the analytical role of the 'labour theory of value' in the determination of the general rate of profit was brought into sharp relief. According to him the explanation of profits in terms of the surplus approach would have been trapped in circular reasoning if the value expression of either aggregate (surplus and capital) were to depend on the rate of profit. The measurement of both aggregates in terms of labour values, which themselves were seen to be independent of distribution, was considered a device to circumvent this danger and provide a non-circular determination of the rate of profit, $r = s/(c + v)$, where r is the general rate of profit, s the 'surplus value', c the value of the means of production or 'constant capital', and v the wages advanced or 'variable capital'. A central message of Marx's *Capital* reads that the rate of profit is positive if and only if there is 'exploitation of workers', i.e. there is a positive 'surplus value'.

In Marx's opinion it was only after the rate of profit had been determined that the problem of normal prices, or 'prices of production' as he called them, could be tackled. Marx dealt with it in terms of a multisectoral analysis of the production of commodities by means of commodities; the deviations of relative prices from labour values are systematically traced back to sectoral differences in the 'organic composition of capital', i.e. the proportion of 'constant' to 'variable' capital (cf. the so-called 'transformation' of values into prices of production; Marx 1959, Part II).

Yet Marx did not fully succeed in overcoming the analytical difficulties encountered by the classical economists in the theory of capital and distribution. He was particularly wrong in assuming that the determination of the rate of profit is logically prior to that of normal prices. Given the system of production and the real wage the rate of profit and prices can be determined only simultaneously. This was first demonstrated by Bortkiewicz (1907). For a rigorous and comprehensive formulation of the classical surplus approach see Sraffa (1960), whose contribution will be dealt with in more detail below.

The Neoclassical Approach

The abandonment of the classical approach and the development of a radically different theory, which came to predominance in the wake of the so-called 'marginalist revolution' in the latter part of the 19th century, was motivated (apart from ideological reasons ever present in debates in capital theory) by the deficiencies of the received (labour) theory of value. Since the new theory was to be an alternative to the classical theory, it

had to be an alternative theory about the same thing, in particular the normal rate of profit. Consequently, the early neoclassical economists, including, for example, Jevons (1871), Walras (1874), Böhm-Bawerk (1889), Wicksell (1893, 1901), and Clark (1899), adopted fundamentally the same method of analysis: the concept of 'long-period equilibrium' is the neoclassical adaptation of the classical concept of normal positions.

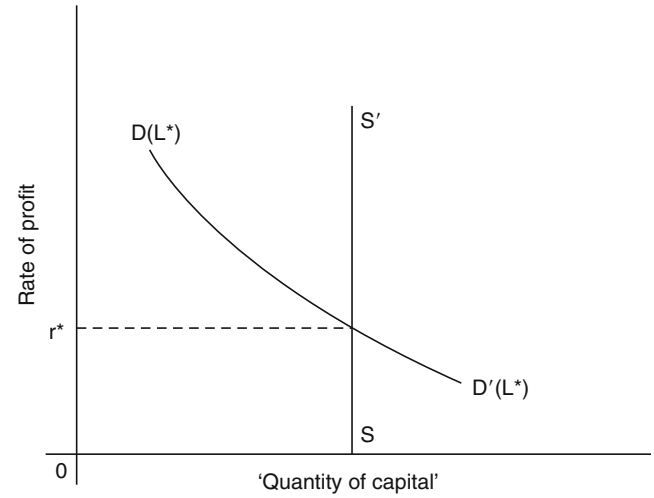
The basic novelty of the new theory consisted in the following. While the surplus approach conceived the real wage as determined *prior* to profits (and rent), in the neoclassical approach all kinds of incomes were explained simultaneously and *symmetrically* in terms of the 'opposing forces' of supply and demand in regard to the services of the respective 'factors of production', labour and 'capital' (and land). It was the seemingly coherent foundation of these notions in terms of *functional* relationships between the price of a service (or good) and the quantity supplied or demanded elaborated by the neoclassical theory that greatly contributed to the latter's success.

As regards the supply side of the neoclassical treatment of capital, careful scrutiny shows that its advocates, with the notable exception of Walras (at least until the fourth edition of the *Eléments*), were well aware of the fact that in order to be consistent with the concept of a long-period equilibrium the capital equipment of the economy could not be conceived as a set of given physical amounts of produced means of production. The 'quantity of capital' in given supply rather had to be expressed in *value* terms, allowing it to assume the physical 'form' best suited to the other data of the theory, i.e. the technical conditions of production and the preferences of agents. For, if the capital endowment is given in kind only a short-period equilibrium, characterized by differential rates of return on the supply prices of the various capital goods, could be established by the forces constituting demand and supply. However, under conditions of free competition, which would enforce a tendency towards a uniform rate of profit, such an equilibrium could not be considered a 'full equilibrium' (Hicks 1932, p. 20).

Thus the formidable problem for the neoclassical approach in attempting the determination of the general rate of profit consisted in the necessity of establishing the notion of a market for 'capital', the quantity of which could be expressed *independently* of the 'price of its service', i.e. the rate of profit. If such a market could be shown to exist, profits could be explained analogously to wages (and other distributive variables) and a theoretical edifice erected on the universal applicability of the principle of demand and supply.

Now, the plausibility of the supply and demand approach to the problem of distribution was felt to hinge upon the demonstration of the existence of a unique and stable equilibrium in the market for 'capital'. (On the importance of uniqueness and stability see, for example, Marshall, 8th edn, 1920, p. 665n.) With the 'quantity of capital' in given supply, this, in turn, implied that a monotonically *decreasing* demand function for capital in terms of the rate of profit had to be established (see Fig. 1). This inverse relationship was arrived at by the neoclassical theorists through the introduction of two kinds of substitutability between 'capital' and labour: substitutability in consumption and in production. According to the former concept a rise in the rate of profit relatively to the wage rate would increase the price of those commodities, whose production is relatively 'capital intensive', compared to those in which relatively little 'capital' per worker is employed. This would generally prompt consumers to shift their demand in favour of a higher proportion of the cheapened commodities, i.e. the 'labour intensive' ones. According to the latter concept a rise in the rate of interest (and thus profits) relatively to wages would make cost-minimizing entrepreneurs in the different industries of the economy employ more of the relatively cheapened factor of production, i.e. labour. Hence, through both routes 'capital' would become substitutable for labour and for any given quantity of labour employed a decreasing demand schedule for capital would obtain. In Fig. 1 the demand schedule DD' corresponding to the *full employment* level of labour L^* (determined simultaneously in the labour market) together with the supply schedule SS' would then ensure a unique and stable equilibrium E with an equilibrium rate of profit r^*. Accordingly, the division of the product between wages and profits

Capital Theory: Debates, Fig. 1



is expressed in terms of the 'scarcity of factors of production', including 'capital' conceived as a value magnitude that is considered independent of the rate of profit.

Let us now briefly look more closely at some of the characteristic features of neoclassical capital theory and point out differences between the main versions in which it was presented.

To define 'capital' as an amount of value requires the specification of the standard of value in which it was to be measured. A rather common procedure was to express capital in terms of consumption goods or, more precisely, to conceive of it as a 'subsistence fund' in support of the 'original' factors of production, labour and land, during the period of production extending from the initial expenditure of the services of these factors to the completion of consumption goods. This notion corresponded to the view that capital resulted from the investment of past savings, which, in turn, implied 'abstention' from consumption. Thus it appeared to be natural to measure 'capital' in terms of some composite unit of consumption goods. However, there was a second dimension of capital contemplated by these authors: the *time* for which capital is invested in a process of production. The idea was that capital can be increased either by using more of it or by lengthening the period of time for which it is invested.

The first author to use time as a single measure of capital was Jevons (1871). The gist of his argument consisted in the concept of a 'production function' $y = f(T)$, where output per unit of labour, y, is 'some continuous function of the time elapsing between the expenditure of labour and the enjoyment of results, T; this function is assumed to exhibit diminishing returns (1871, pp. 240–41). Jevons showed that in equilibrium $r = f'(T)/f(T)$.

Jevons's contribution was the starting point of the Austrian theory of capital and interest with Böhm-Bawerk and Wicksell as its main representatives. Böhm-Bawerk's concern was with establishing a temporal version of the demand and supply approach. This involved the appropriate reformulation of the data of the theory. The central elements of his analysis were the concepts of 'time preference' and the 'average period of production', used in describing consumer preferences and technical alternatives, respectively. As in Jevons social capital was conceived as a subsistence fund and was seen to permit the adoption of more productive but also more 'roundabout', i.e. time-consuming, methods of production. It was to the concept of the 'average period of production' that the marginal productivity condition was applied in the determination of the rate of interest.

Among the older neoclassical economists it was perhaps Wicksell who understood best the difficulties related to the problem of a unified treatment of capital in terms of the demand and supply approach. In particular, Wicksell was

critical of attempts to work with the value of capital as a factor of production alongside the physically specified factors of labour and land in the production function of single commodities. This implied 'arguing in a circle' ([1901] 1934, p. 149), since capital and the rate of interest enter as a cost in the production of capital goods themselves. Hence the value of the capital goods inserted in the production function depends on the rate of interest and will change with it. Moreover, Wicksell expressed doubts as to the possibility of providing a sufficiently general definition of the 'average period of production' that could be used to represent capital in a way that is not threatened by this kind of circularity. In the *Lectures* he tried to overcome these difficulties by introducing production functions in terms of *dated* services of the 'original' factors labour and land.

While Wicksell shared Böhm-Bawerk's procedure of conceiving the 'capital endowment' of the economy as a value magnitude, he become increasingly sceptical whether it was admissible to identify it with some unspecified stock of subsistence goods, which, in turn, was seen to provide some measure of 'real' capital. With capital as a value magnitude Wicksell showed that the rate of interest is generally not equal to the marginal productivity of 'capital'. This discrepancy is due to the revaluation of the capital stock entailed by a change in distribution. The phenomenon is known as the 'Wicksell effect' and was regarded by Joan Robinson as the key to a criticism of the marginal productivity theory of income distribution.

Authors like J.B. Clark and Marshall appear to have been less aware of the fact that the conditions of production of single commodities cannot be defined in terms of production functions that include 'capital' among the factors of production. Obviously, the criticism levelled against these versions applies also to the concept of the 'aggregate production function', which boomed in the late 1950s and throughout the 1960s in conjunction with neoclassical growth theory.

Alternative views of the fundamentals of capital theory were expressed in a controversy between Böhm-Bawerk and J.B. Clark around the turn of this century (cf. in particular Böhm-Bawerk 1906–7; Clark 1907). Böhm-Bawerk criticized Clark's attempt to differentiate between 'true capital', a permanent abiding fund of productive wealth, and 'concrete capital goods', each of which is destructible and has to be destroyed in order to serve its productive purpose; in Böhm-Bawerk's view this is 'dark, mystical rhetoric'. Furthermore, Böhm-Bawerk refuted Clark's claim that no concept of 'waiting' or 'abstinence' is needed to explain interest in stationary equilibrium. Without some concept of time preference, and thus a theory of saving, the determination of the rate of interest is left hanging in the air.

Irving Fisher (1930) extended general equilibrium theory to intertemporal choices. However, he proceeded as if there were a single composite commodity to be produced and consumed at different dates. In his discussion of the theory of interest all prices, wages and rents are assumed to be fixed. Hence the interrelationship between the rate of interest, prices and the remaining distribution variables is set aside. The 'investment opportunities' available to an individual and to society as a whole are summarized in intertemporal production possibility frontiers. Due to the assumption of diminishing returns Fisher arrived at a decreasing demand function for saving with respect to the rate of interest. As Keynes noted, this is equivalent to his 'marginal efficiency of capital' schedule (Keynes 1936, p. 140). Because of 'impatience' the supply of saving is considered to be positively related to the rate of interest. The market equilibrium between the supply of, and the demand for, saving gives the rate of interest, which is equal to the marginal rate of return over the cost of the marginal increase in the capital stock. (For an attempt to generalize Fisher's rate of return approach see Solow 1967. For a critique of Fisher and Solow see Pasinetti 1969; Eatwell 1976).

The 1930s brought a further controversy on the theory of capital (cf. Kaldor 1937). This was triggered off by a series of articles by F.H. Knight (e.g. Knight 1934), in which he launched an attack on the concept of the 'period of production' revived a few years earlier by Hayek, among others. In particular, Knight argued that there is no need to refer to a 'quantity of capital' and that therefore the

'vicious circle' disappears. The rate of interest could be ascertained with reference to the instantaneous rate of investment and the present value of the additional stream of future income generated by it, However, Knight's proposed solution to the problem of circularity in terms of a 'theory of capital without capital' is illusory, since if the accusation of circularity applies at all (because the value of capital goods cannot be ascertained independently of the rate of interest), it applies both to the stock variable 'capital' and the corresponding flow variable 'investment'.

Finally, some recent attempts to revive and reformulate basic elements of the doctrines of the older neoclassical and Austrian authors should be noted, in particular: Weizsäcker (1971), Hicks (1973), and Faber (1979) on the Austrian theory, Morishima (1977) on Walras, and Hirshleifer (1970) and Dougherty (1980) on Fisher. (For a critical assessmenty of the older theories see especially Garegnani 1960).

The Recent Critique of Neoclassical Theory

Sraffa (1960) deserves the credit for having elaborated a consistent formulation of the classical surplus approach to the problem of capital and distribution. His analysis provided the fundamental basis for a critique of the prevalent neoclassical theory during the so-called 'Cambridge controversies in the theory of capital' (see Harcourt 1969; Kurz 1985).

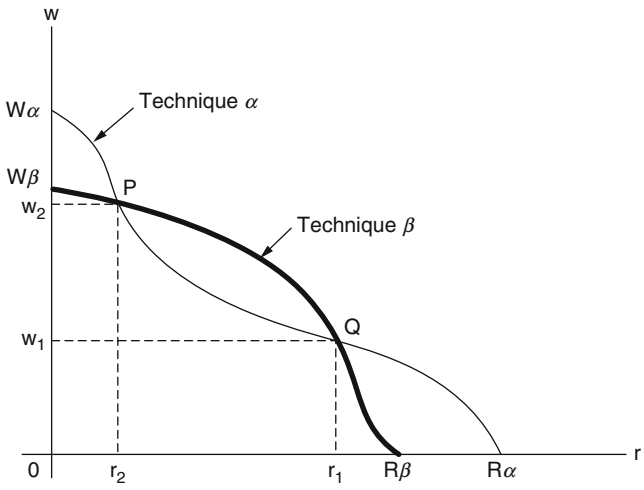
Sraffa starts from a given system of production in use in which commodities are produced by means of commodities. If wages are assumed to be paid at the end of the uniform production period, then, in the case of single-product industries (i.e. circulating capital only) and with gross outputs of the different products all measured in physical terms and made equal to unity by choice of units, we have the price system

$$p = (1 + r)Ap + wl,$$

where p is the column vector of normal prices, A is the square matrix of material inputs, l is the vector of direct labour inputs and w is the wage rate. Under certain economically meaningful conditions, for any given feasible wage rate in terms of a given standard the above equation yields a unique and strictly positive price vector in terms of the standard and a unique and non-negative value of the rate of profit. The investigation of the 'effects' of variations in one of the distribution variables on the other one and on the prices of commodities, assuming that the methods of production remain unchanged, yields the following results. First, the system possesses a finite maximum rate of profits $R > 0$ corresponding to a zero wage rate. Second, the vector of prices in terms of the wage rate p/w (prices in terms of quantities of *labour commanded*) is positive and rises monotonically for $0 \leq r < R$, tending to infinity as r approaches R. Third, at the maximum level of wages corresponding to $r = 0$ relative prices are in proportion to their labour costs, while at $r > 0$ relative prices generally deviate from relative labour costs and vary with changes in r (or w); it is only in the special case of uniform 'proportions' of labour to means of production in all industries that prices are proportional to 'labour values' for all levels of r (w). (For a discussion of joint production, fixed capital and land, see Pasinetti 1980.)

While earlier authors were of the opinion that the capital-labour or capital-output ratios of the different industries could be brought into a ranking that is independent of distribution, this is generally not possible: 'the price of a product. . .may rise or it may fall, or it may even alternate in rising and falling, relative to its means of production' (Sraffa 1960, p. 15). This result destroys the foundation of those versions of the traditional theory that attempted to define the conditions of production in terms of production functions with 'capital' as a factor. Moreover, as regards the concept of the 'capital endowment' of the economy conceived as a value magnitude, the same 'real' capital may assume different values depending on the level of r. Sraffa concludes that these findings 'cannot be reconciled with *any* notion of capital as a measurable quantity independent of distribution and prices' (1960, p. 38).

Capital Theory: Debates, Fig. 2

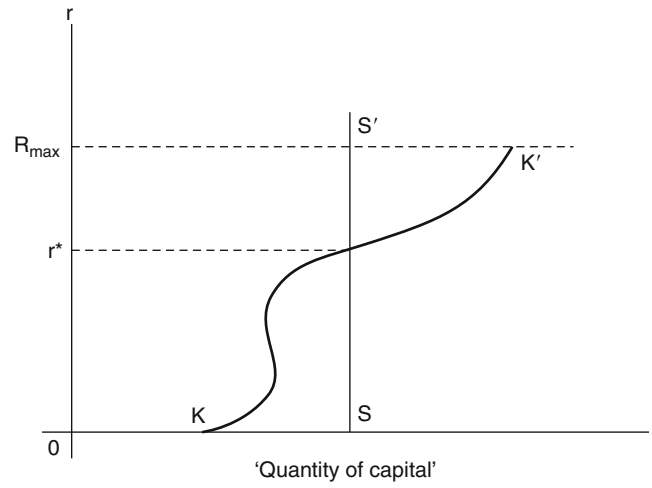


Samuelson (1962), in an attempt to counter Joan Robinson's (1953) attack on the aggregate production function, claimed that even in cases with heterogeneous capital goods some rationalization can be provided for the validity of simple neoclassical 'parables' which assume there is a single homogeneous factor called 'capital', the marginal product of which equals the rate of interest. But, alas, Samuelson based his defence of traditional theory in terms of the construction of a 'surrogate production function' on the assumption of equal input proportions (cf. 1962, pp. 196–7). By this token the 'real' economy with heterogeneous goods was turned into the 'imaginary' economy with a homogeneous output, i.e. the 'surrogate production function' was nothing more than the infamous aggregate production function. (For a critique of Samuelson's approach see particularly Garegnani 1970).

Implicit in the above system of price equations is the inverse relationship between the wage and the rate of profit, or *wage curve*, of the given system of production, $w = w(r)$. We may now turn to the hypothesis that for one or several industries alternative technical methods are available for the production of the corresponding commodity. The technology of the economic system as a whole will then be represented by a series of alternative techniques obtained from all the possible combinations of methods of production for the various commodities. Expressing w and p in terms of a commodity produced in all the alternative systems, we obtain as many different wage curves as there are alternative techniques. In Fig. 2 it is assumed that only two techniques, α and β exist. Clearly, at any level of the wage rate (or rate of profit), enterpreneurs will choose the *cost-minimizing* system of production. It can be shown that, whichever the system initially in use, the tendency of producers to switch to the cheaper system will bring them to the one giving the highest rate of profit (wage rate), whereas systems giving the same r for the same w will be indifferent and can coexist. Thus, in the example of Fig. 2, in the two intervals $0 < w < w_1$ and $w_2 < w \leq W_\alpha$ technique α will be chosen, while in the interval $w_1 < w \leq W_2$ technique β turns out to be superior; at the two switch points P and Q both techniques are equiprofitable. It follows that with a choice of technique the relationship between w and r, or *wage frontier*, will be represented by the outermost segments or envelope of the intersecting wage curves.

Figure 2 shows that the same technique (α) may be the most profitable of a number of techniques at more than one level of the wage rate even though other techniques (here β) are more profitable at wage rates in between. The implication of this possibility of the *reswitching* of techniques is that the direction of change of the input proportions cannot be related unambiguously to

Capital Theory: Debates, Fig. 3



changes of the so-called 'factor prices'. The central element of the neoclassical explanation of distribution in terms of supply and demand is thus revealed as defective. This element consisted in the proposition that a rise of r must decrease the 'quantity of capital' relative to labour in the production of a commodity because of the assumed substitutability in production and consumption. The demonstration that a rise in r may lead to the adoption of the more 'capital intensive' of two techniques clearly destroys the neoclassical concept of substitution in production. Moreover, since a rise in r may cheapen some of the commodities, the production of which at a lower level of r was characterized by a relatively high 'capital intensity', the substitution among consumption goods contemplated by the traditional theory of consumer demand may result in a higher, as well as in a lower, 'capital intensity'. It follows that the principle of substitution in consumption cannot offset the breakdown of the principle of substitution in production. Finally, it is worth mentioning that reswitching is not necessary for *capital-reversing* cf. Symposium 1966, p. 516).

The negative implication of reverse capital deepening for traditional theory can be illustrated by means of the example of Fig. 3, in which the value of capital corresponding to the full employment level of labour is plotted against the rate of profit. Obviously, if with traditional analysis we conceived the curve KK' as the 'demand curve' for capital, which, together with the corresponding 'supply curve' SS', is taken to determine the 'equilibrium' level of r, we would have to conclude that this equilibrium, although unique, is unstable. With free competition and perfectly flexible distributive variables a deviation of r from r^* would lead to the complete extinction of one of the two income categories. According to the critics of traditional theory, the finding that the quantity of a factor demanded need not be related to the price of the factor service in the conventional, inverse manner demonstrates the failure of the supply and demand approach to the explanation of normal distribution, prices and quantities.

Neoclassical Responses

Neoclassical economists tried to counter the attack in various ways. At first it was claimed that reswitching is impossible. When this claim was shown conclusively to be false (cf. Symposium 1966), doubts were raised as to its empirical importance (see, for example, Ferguson 1969), thereby insinuating that neoclassical theory was a simplified picture of reality, the basic correctness of which would not be endangered by 'exceptions' of the kind analysed in the capital debate. Other

advocates of the neoclassical approach were conscious of how defective the attempt was to play down the importance of reswitching and capital-reversing using the 'empirical' route. Since the phenomenon was irrefutable it had to be absorbed and shown to be compatible with the more sophisticated versions of the dominant theory.

Perhaps the first move in this direction was made by Bruno et al. (1966), who drew an analogy between reswitching and the long-known possibility of the existence of multiple internal rates of return. However, whereas the latter phenomenon is a discovery within the partial, 'fixed-price' framework of microeconomic theory of investment, reswitching presupposes a total, general framework. Moreover, we are not told how traditional theory was both able to cope with reswitching and yet preserve its basic structure.

A more interesting challenge came from authors such as Bliss (1975) and Hahn (1982). They contended that because of its concern with a uniform rate of profit Sraffa's analysis can be considered a 'special case' of general equilibrium theory. According to these authors the criticism of traditional neoclassical capital theory implicit in Sraffa is correct but has no bearing upon modern general equilibrium theory. Since in the latter the distribution of income is explained in terms of given *physical* endowments of agents, there is no need to find a scalar representation of the capital stock. The uniformity of profit rates is taken to be 'a very special state of the economy' (Hahn 1982, p. 363) which, for given preferences and production sets, presupposes a particular composition of initial endowments. In general, there will be as many own rates of return as there are different assets in the endowment set.

The first thing to be noticed is that the preservation of the basic supply and demand approach to the explanation of prices, distribution and quantities in modern general equilibrium theory is effectuated at the cost of the abandonment of the traditional long-period method. As we have seen, this method was shared by all 'forerunners' of this theory, including, most notably, Walras and von Neumann (1936). Indeed, the change in the notion of equilibrium involved expresses a fundamental break with the analytical method used by all economic theory up to the 1930s, when partly because of a growing perception among neoclassical economists that the whole approach was threatened by the difficulties concerning the notion of capital a drastic methodological reorientation was advocated (cf. Garegnani 1976; Milgate 1979). Most influential in this move away from the traditional method was apparently Hicks's *Value and Capital* (1939; second edition 1946). Interestingly enough, Hicks himself appears to have become increasingly sceptical as to the usefulness of the 'temporary equilibrium method' then suggested by him (see, for example, Hicks 1965, pp. 73–4).

The second observation concerns Hahn's attempt to interpret Sraffa's analysis as a special case of general equilibrium theory. Since the latter takes as data (i) the preferences of consumers, (ii) the technical conditions of production, and (iii) the physical endowments, Hahn's view necessarily leads to the question of which constellation of these data is compatible with a uniform rate of profit. Clearly, to superimpose the latter specification on an ordinary general equilibrium system would render it over-determined, as some of the older neoclassical authors were well aware of. Hence, following the interpretation under consideration, (i), (ii) or (iii) cannot be taken as independent variables. Now it is Hahn's contention that at the basis of Sraffa's price equations there must be a special proportion between the initial endowments; i.e. (iii) is tacitly assumed to be specified accordingly. However, as we have seen there is no evidence in support of this presupposition. The surplus approach does not require given endowments of produced means of production in order to determine distribution and normal prices. In fact, looking at classical analysis as a whole the quantities of the capital goods available may be considered as dependent rather than independent variables. In analysing the problem of value, capital and distribution the classical economists took the capital stocks installed in the different industries as exactly adjusted to *given outputs*, such that the latter could be produced at minimum costs. The tendency towards normal capital utilization and a uniform rate of profit was seen to be the outcome of the working of

the persistent forces of the system reflected in the competitive decisions of producers.

Since the opinion entertained by Hahn that Sraffa's analysis can be subsumed as a 'special case' under modern neoclassical theory has to be rejected, the question remains, which of the two is the more powerful instrument of analysis. There does not seem to exist a ready-made answer at present. The following remarks on the dominant neoclassical theory must suffice.

Obviously, to take the capital endowment as given in kind implies that only 'short-period' equilibria can be determined. Because firms 'prefer more profit to less' (Hahn 1982, p. 354) the size and composition of the capital stock will rapidly change. Thus, major factors which general equilibrium theory envisages as determining prices and quantities are themselves subject to quick changes. This, in turn, makes it difficult to distinguish them from those accidental and temporary factors, which, at any given moment of time, prevent the economy from settling in the position of equilibrium. More important, the fast variation in relative prices necessitates the consideration of the influence of future states of the world on the present situation.

This can be approached in two different ways. First, if there were complete futures markets the analysis could be carried out in terms of the concept of *intertemporal equilibrium*. However, the assumption that all intertemporal and all contingent markets exist, which has the effect of collapsing the future into the present, can be rejected on grounds of realism and economic reasoning (see, for example, Bliss 1975, pp. 48 and 61). Moreover, there is the following conceptual problem (see Schefold 1985). If in equilibrium some of the capital stocks turn out to be in excess supply these stocks assume zero prices. This possibility appears to indicate that the expectations entrepreneurs held in the past when deciding to build up the present capital stocks are not realized. Hence, strictly speaking we are faced with a disequilibrium situation because otherwise the wrong stocks could not have accumulated. Therefore, the problem arises how the past or, more exactly, possible discrepancies between expectations and facts influence the future.

Since the notion of intertemporal equilibrium cannot be sustained the theory is ultimately referred back to the introduction of individual price expectations concerning future deliveries of commodities for which no present markets exist. This leads to the *temporary equilibrium* version of modern neoclassical theory. The basic weakness of the theories of temporary equilibrium concerns the necessarily arbitrary choice of hypotheses about individual price expectations. Indeed, as Burmeister stresses, 'all too often "nearly anything can happen" is the only possible unaqualified conclusion' (Burmeister 1980, p. 215). Moreover, the stability properties of this kind of equilibrium are unclear, since small perturbations caused by accidental factors may entail changes in expectations, which define that very equilibrium.

The danger of lapsing into empty formalism and of depriving the theory of clear-cut results was of course recognized by several supply and demand theorists and considered a fundamental weakness. In view of it some of them were prepared to dispence with the alleged generality of general equilibrium theory and return to some version of traditional neoclassical analysis. After the recent debate in capital theory this involved ruling out reswitching and other 'perverse', i.e. non-conventional, phenomena in terms of sufficiently bold assumptions about available techniques. It comes as no surprise that given these assumptions the central neoclassical postulate of the inverse relation between the capital-labour ratio and the rate of profit should re-emerge as 'one of the most powerful theorems in economic theory' (Sato 1974, p. 355). However, in order to be clear about this move it deserves to be stressed that it was motivated, as one author expressly admits, by the fact that 'regular economies' have 'desirable properties' (Burmeister 1980, p. 124).

See Also

- ▶ Accumulation of Capital
- ▶ Marginal Productivity Theory
- ▶ Reverse Capital Deepening

Bibliography

Bliss, C.J. 1975. *Capital theory and the distribution of income*. Amsterdam: North-Holland.

Bruno, M., E. Burmeister, and E. Sheshinski. 1966. The nature and implications of the reswitching of techniques. *Quarterly Journal of Economics* 80: 526–553.

Burmeister, E. 1980. *Capital theory and dynamics*. Cambridge: Cambridge University Press.

Clark, J.B. 1899. *The distribution of wealth*. London: Macmillan.

Clark, J.B. 1907. Concerning the nature of capital: A reply. *Quarterly Journal of Economics* 21: 351–370.

Dougherty, C. 1980. *Interest and profit*. London: Methuen.

Eatwell, J. 1976. Irving Fisher's 'Rate of return over cost' and the rate of profit in a capitalistic economy. In *Essays in modern capital theory*, ed. M. Brown, K. Sato, and P. Zarembka. Amsterdam: North-Holland.

Faber, M. 1979. *Introduction to modern Austrian Capital Theory*. Berlin: Springer.

Ferguson, C.E. 1969. *The neoclassical theory of production and distribution*. Cambridge: Cambridge University Press.

Fisher, I. 1930. *The theory of interest*. London: Macmillan.

Garegnani, P. 1960. *Il capitale nelle teorie della distribuzione*. Milan: Giuffrè.

Garegnani, P. 1970. Heterogeneous capital, the production function and the theory of distribution. *Review of Economic Studies* 37(3): 407–436.

Garegnani, P. 1976. On a change in the notion of equilibrium in recent work on value. In *Essays in modern capital theory*, ed. M. Brown, K. Sato, and P. Zarembka. Amsterdam: North-Holland.

Hahn, F.H. 1982. The neo-Ricardians. *Cambridge Journal of Economics* 6(4): 353–374.

Harcourt, G.C. 1969. Some Cambridge controversies in the theory of capital. *Journal of Economic Literature* 7(2): 369–405.

Hicks, J.R. 1932. *The theory of wages*. London: Macmillan.

Hicks, J.R. 1939. *Value and capital*, 2nd ed, 1946. Oxford: Clarendon Press.

Hicks, J.R. 1965. *Capital and growth*. Oxford: Oxford University Press.

Hicks, J.R. 1973. *Capital and time – A neo-Austrian theory*. Oxford: Oxford University Press.

Hirshleifer, J. 1970. *Investment, interest and capital*. Englewood Cliffs: Prentice-Hall.

Jevons, W.S. 1871. *The theory of political economy*. New York: Kelley. Reprint.

Kaldor, N. 1937. The recent controversy on the theory of capital. *Econometrica* 5: 201–233.

Keynes, J.M. 1936. *The general theory of employment, interest and money*. London: Macmillan.

Knight, F.H. 1921. *Risk, uncertainty and profit*. Chicago: University of Chicago Press.

Knight, F.H. 1934. Capital, time and the interest rate. *Economica* 1: 257–286.

Kurz, H.D. 1985. Sraffa's contribution to the debate in capital theory. *Contributions to Political Economy* 4: 3–24.

Marshall, A. 1890. *Principles of economics*, 8th edn (1920). Reprint, reset, London: Macmillan, 1977.

Marx, K. 1894. *Capital*, vol. III. Moscow: Progress Publishers; Harmondsworth: Penguin, 1959.

Milgate, M. 1979. On the origin of the notion of 'intertemporal equilibrium'. *Economica* 46: 1–10.

Morishima, M. 1977. *Walras' economics: A pure theory of capital and money*. Cambridge: Cambridge University Press.

Pasinetti, L.L. 1969. Switches of technique and the 'rate of return' in capital theory. *Economic Journal* 79: 508–531.

Pasinetti, L.L. (ed.). 1980. *Essays on the theory of joint production*. London: Macmillan.

Ricardo, D. 1951–73. *The works and correspondence of David Ricardo*. 11 vols, ed. P. Sraffa in collaboration with M.H. Dobb, Cambridge: Cambridge University Press.

Robinson, J. 1953. The production function and the theory of capital. *Review of Economic Studies* 21(2): 81–106.

Samuelson, P.A. 1962. Parable and realism in capital theory: The surrogate production function. *Review of Economic Studies* 29: 193–206.

Sato, K. 1974. The neoclassical postulate and the technology frontier in capital theory. *Quarterly Journal of Economics* 88(3): 353–384.

Schefold, B. 1985. Cambridge price theory: Special model or general theory of value? *American Economic Review: Papers and Proceedings* 75(2): 140–145.

Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations*. Ed. E. Cannan, introduced by G.J. Stigler, Chicago: Chicago University Press, 1976.

Solow, R.M. 1967. The interest rate and the transition between techniques. In *Socialism, capitalism and economic growth, essays presented to Maurice Dobb*, ed. C.H. Feinstein. Cambridge: Cambridge University Press.

Sraffa, P. 1951. Introduction. In *The works and correspondence of David Ricardo*, vol. 1, ed. D. Ricardo.

Sraffa, P. 1960. *Production of commodities by means of commodities*. Cambridge: Cambridge University Press.

Symposium. 1966. On paradoxes in capital theory: a symposium. *Quarterly Journal of Economics* 80(4): 526–583.

von Böhm-Bawerk, E. 1889. *Positive Theorie des Kapitales*. Jena: Gustav Fischer. Trans. as *Positive theory of capital*. London: Smart, 1891.

von Böhm-Bawerk, E. 1906–7. Capital and interest once more. *Quarterly Journal of Economics* 21(Pt. I), Nov 1906, 1–21(Pt. II), Apr 1907, 247–282.

von Bortkiewicz, L. 1907. Zur Berichtigung der grundlegenden theoretischen Konstruktion von Marx im dritten Bande des 'Kapital'. *Jahrbücher für Nationalökonomie und Statistik*, Jul. English trans. in Appendix to E. von. Böhm-Bawerk, *Karl Marx and the Close of his System*, ed. P. Sweezy, New York, 1949.

von Neumann, J. 1936. Über ein ökonomisches Gleichungssystem und eine Verallgemeinerung des Browerschen Fixpunktsatzes. In: *Ergebnisse eines Mathematischen Kolloquiums*, ed. K. Menger, Vienna: F. Deuticke. Trans. as 'A Model of General Economic equilibrium', *Review of Economic Studies* 13(1), (1945–6), Winter, 1–9.
von Weizsäcker, C. Ch. 1971. *Steady state capital theory.* Berlin: Springer.
Walras, L. 1874–7. *Eléments d'économie politique pure*. Lausanne: Corbaz. 4th edn 1900. Trans. by W. Jaffé of definitive edn (1926) as *Elements of pure economics*. London: Allen & Unwin, 1954.
Wicksell, K. 1893.*Über Wert, Kapital und Rente*. Jena: Gustav Fischer. Trans. as *value, capital and rent*. New York: Kelley, 1954.
Wicksell, K. 1901. *Föreläsingar i Nationalekonomi*, vol. 1. Lund: Berlingska Boktryckeriet. Trans. as *Lectures on political economy*, vol. 1. London: Routledge & Kegan Paul, 1934.

Capital Utilization

Roger Betancourt

Abstract

Utilization of capital can take place through variations in the duration of working time, given intensity, or through variations in the intensity of working time, given duration, or both. This article focuses on the economic factors determining duration and discusses the issues affecting and affected by variations in intensity. The latter can take the form of variations in speed or in the use of inputs that are complements to capital relative to some maximum or optimum. We provide a historical perspective, discuss modern theory, its main applications and links to the issues of speed and capacity, and identify important implications.

Keywords

Agency costs; Business cycles; Capacity utilization; Capital utilization; Compensating differentials; Depreciation; Duality; Duration; Dynamic factor demand models; Elasticity of substitution; Exploitation; General equilibrium; Incomplete contracts; Leontief production function; Market imperfections; Marshall, A.; Marx, K. H.; Monitoring; Non-convexity; Obsolescence; Overtime; Partial equilibrium; Production function; Shift work; Shift differential; Solow residual; Speed; Specific-factors models; Taxation of corporate profits; Technical change; Transaction costs; Two-sector models; Wear and tear; Work day of capital; Work day of labour

JEL Classifications

D2; D24; E22; E23; L23; J81

Capital utilization is given different interpretations in the economic literature. If a machine is available for use during, say, a day, then various levels of utilization can be obtained by varying the duration of operations within the day. For any fixed duration within the day, however, it is also possible to vary the machine's rate of utilization by varying its speed. In each case there is variation in capital utilization, but both physical and economic characteristics differ widely in the two cases. Moreover, even with duration and speed constant within the day, some writers define variations in capacity utilization via variations in the variable inputs employed with a given machine per day relative to some maximum or optimum daily output. Unfortunately, these as well as other writers frequently use the terms 'capital utilization' and 'capacity utilization' interchangeably.

The discussion here will focus on the analysis of variations in the duration of operations. A brief historical perspective sets the stage for a presentation of modern theory and applications, including links to the issues of speed and capacity. A succinct conclusion provides implications for closely related economic issues.

Historical Perspective

Concern with the duration of operations dates to the late 18th century and the spread of the factory system in England. Early writing emphasized the appropriate length of the working day relative to its social consequence for workers and its economic

consequence for capitalists. Positions on these issues were developed in the context of debates over the various Factory Acts in England. These discussions usually assumed the length of the working day to be the same for capital and labour.

Marx provides a most interesting example of the development of economic thinking on duration up to his time. The length of the working day is given substantial attention in his work (1867, ch. 10); indeed, it provides the cornerstone for his theory of exploitation (see, for example, Morishima 1973, ch. 5); yet Marx pays only minor attention to the separation of capital's work day from labour's work day which is at the centre of modern analysis.

Marshall, like his predecessors, was interested in duration because of its implications for the well-being of workers and the viability of the economic system. But he saw the separation of the work day of labour from the work day of capital inherent in shift-work systems as an opportunity for resolving the conflicting interests of workers and capitalists with respect to the length of the work day. Thus he becomes an advocate of the adoption of multiple shifts early in his professional career (1873) and maintained his interest in the topic throughout his career (see, for example, 1923, p. 650).

Marshall's emphasis became the basis for the work of Robin Marris (1964), who treats capital utilization as a synonym for shift-work. Interestingly enough, the other modern pioneer, Georgescu-Roegen (for example, 1972), stresses the choice of the daily duration of operations, acknowledges Marx's emphasis on the topic, but overlooks Marshall as well as Marris. Both view the choice of duration at the plant level, either directly or through the selection of a shift-work system, as a long-run or *ex ante* decision, that is, before the plant is built. Moreover, both assume the *ex post* elasticity of substitution to be zero, that is, within the day no variations in choice of technique are allowed once the factory is built. However, while Marris uses discrete techniques of production and discrete systems of utilization to describe the structure of the firm's optimization problem, Georgescu-Roegen uses a continuous production function and a continuous index of the daily duration of operations; these differences of method do not generate substantial differences in results.

Both economists use their analyses to argue against anachronistic social legislation and draw implications from their work for an important contemporary economic problem, namely, the improvement of economic conditions in developing countries.

Before presenting the modern theory and its applications it is useful to note a few salient facts. Thanks to Foss's efforts (1981) there are reliable estimates of the average workweek of capital (plant hours) in US manufacturing for 1929 and 1976–67 and 82 hours, respectively. These estimates can be compared to an average workweek for labour of 50 hours in 1929 and 40 hours in 1976. Furthermore, Foss views the rise in capital's workweek between 1929 and 1976 as an underestimate of the increase in shift-work, because of the decrease in the number of days worked per week during this same period. The most thorough update of this data work is Beaulieu and Mattey (1998). It generates an average workweek of capital for manufacturing during the period 1974–92 of 97 hours per week. These 'facts' underlie interest in the topic and the frequent identification of capital utilization with shift-work.

Modern Theory and Applications

A number of contributions have incorporated the choice of duration into the neoclassical theory of the firm. This work is most concisely exposited using a model which relies on duality theory to generate the main results available in this literature (see Betancourt 1986).

The firm's optimization problem is viewed as a two-stage procedure. In the first stage the decision-maker generates a cost function for each given level of duration; in the second stage the decision-maker selects from these cost functions that one which leads to least total cost. The end result in the two-input case is:

$$C^* = dC\left(w^*, r^*, x^*\right). \tag{1}$$

For a given reference unit of duration, w^* represents the average wage rate, r^* the price of

capital services, x^* the level of output, while d represents an index of duration of operations, C is a classical cost function, and C^* represents the total cost of operations at the optimal level of duration.

For example, if an eight-hour shift starting during normal hours is the reference unit of duration, as duration increases beyond this reference period: the average wage rate (w^*) increases because of shift differentials due to workers' preferences for normal hours or social legislation; and the price of capital services per eight-hour shift decreases, although there will be two opposite tendencies in this case. The daily price of a unit of capital increases due to the additional wear and tear created by the longer duration, but this price is now spread over a greater number of hours, and the price of capital services per eight-hour shift (r^*) decreases. Betancourt and Clague (1981, ch. 2, sect. 2) provide a detailed discussion of why the second effect predominates. Finally, as duration increases, the same daily output is spread over a greater number of hours, and the level of output per eight-hour shift (x^*) decreases.

The formulation in (1) yields the main insights about capital utilization or shiftwork at the plant level offered by the early literature that followed Georgescu- Roegen and Marris. A brief listing of these results is as follows: (i) high shift differentials or overtime rates discourage capital utilization by increasing w^*; (ii) technologies with high degrees of returns to scale discourage utilization by raising the costs of operating at low levels of output (x^*); (iii) technologies with high degrees of capital intensity encourage capital utilization because the consequent fall in the relevant cost of capital (r^*) affects a higher percentage of costs; and (iv) technologies with abundant *ex ante* substitution possibilities encourage utilization because they lower the costs of taking advantage of the consequent fall in the cost of capital (r^*) through the building of a more capital intensive factory. These four factors are the main long-run determinants of optimal duration on the cost side.

In addition, two other characteristics of the utilization decision are worth stating. First, factories built to operate at high levels of utilization will be designed to use capital-intensive techniques. Second, how exogenous changes in input costs affect duration depends critically on the *ex ante* elasticity of substitution. For instance, if this elasticity is greater than unity, under constant returns to scale an exogenous fall in the price of capital lowers the costs of building the plant to operate longer hours.

One application of the model is as the theoretical basis for empirical studies of the choice of duration at the plant level. The model's implications were consistent with several different bodies of plant level data (see Betancourt and Clague 1981, chs 4–8) across non-continuous process industries. Recent work using more detailed plant level data for specific industries, for example automobiles, confirms the role of the number of shifts as a long-run margin of adjustment and it stresses the importance of changes in duration through overtime and daily closings as short-run margins of adjustment in the United States (Bresnahan and Ramey 1994). Detailed studies of the auto industry for Europe and Japan (Anxo et al. 1995, chs 12 and 13, respectively) are also consistent with this long-run role for the number of shifts. Mayshar and Halevy (1997) develop a model that allows for *ex post* substitution possibilities as a short-run margin of adjustment. The above studies imply that there is a choice of duration, even in the short run, but in some industries continuous processes dominate and the choice is really to operate or not operate the process. A major extension of the model that captures this feature is provided by Das (1992), who develops and estimates a discrete dynamic programming model for the cement industry at the kiln level. In this context a plant is basically an additive collection of kilns and Das allows for three decisions, namely, operate, retire or keep idle a kiln in any plant.

Alternative approaches to the non-convexities that arise at the plant level have been developed by looking at the industry as the unit of analysis. Prucha and Nadiri (1996) provide an insightful and sophisticated example of this option applied to the US electrical machinery industry by making endogenous the capital utilization decision in the

context of dynamic factor demand models. In a similar industry setting, Cardellichio (1990) uses the assumption of Leontief production functions at the mill level to analyse utilization for the lumber industry as a whole.

From a theoretical perspective an application of the model in (1) has been as the basis for the choice of duration in standard two-sector general equilibrium models. In the context of the international trade literature, Betancourt et al. (1985), for example, use the specific-factors model with variable utilization to reconcile the dual scarcity explanation of Anglo-American trade in the 19th century with the empirical evidence on observed utilization levels. In the context of the public finance literature Coates (1991) generalizes the standard analysis of the incidence of the corporate profits tax by allowing for variable utilization. He concludes that overestimates of the burden of the tax in the order of 10–60 per cent are most probable as a result of ignoring this long-run margin of adjustment in a general equilibrium context. A more abstract general equilibrium approach allowing for firm's decisions over duration and starting times as well as for worker's preferences over these work schedules has been developed recently by Garcia Sanchez and Vazquez Mendez (2005). Its main substantive result replicates one partial equilibrium result noted above, namely, that high capital intensity in the form of a high capital-labour ratio leads to an increase in utilization.

A short-run perspective has played an important role in dramatizing the policy implications of high levels of utilization for employment and output, since in this perspective a doubling of utilization implies a doubling of employment and output. Nevertheless a long-run perspective (see Betancourt and Clague 1981, chs 9–11) provides a far less optimistic view about the likelihood of these outcomes. Ironically the evaluation of a shorter workweek for labour in Europe, which is analytically similar, has been carried out primarily from a short-run perspective (for example, Anxo et al. 1995, ch. 14). Garcia Sanchez and Vazquez Mendez (2005), however, suggest this topic as one for potential application of their long-run model.

Related Issues: Speed and Capacity

The relations between duration, speed and capacity are difficult to analyse and provide an opportunity for confusion. To start, consider a dual representation of the cost function in (1). Namely,

$$x = dF(K, L) \tag{2}$$

where x is the level of daily output, that is, $x = \mathrm{d}x^* = \mathrm{d}F$; F is a neoclassical production function defined over the reference period of duration; K represents both the level of the capital stock employed and the rate of capital services, which implies that the speed of operations (v) is constant and set at unity; and L represents labour services per reference period of duration. Alternatively, those who analyse variations in utilization through choice of speed represent the productive process as follows:

$$x = F(vK, L) \tag{3}$$

where all variables have been previously defined. In (3) duration is set at unity.

Writers who employ (3) assume that the price of the capital stock is an increasing function of speed or utilization (for example, Smith 1970). Since costs are defined as

$C = r(v)K + wL$, where $r'(v) > 0$, the cost of a unit of capital services obtained by increasing speed is an increasing function of v. While in the duration model the price of the capital stock $r(d)$ is an increasing function of duration ($r'(d) > 0$), the cost of a unit of capital services obtained by increasing duration is a decreasing function of duration, that is, $r^* = r(d)/d$ and $r^{*\prime}(d) < 0$. This difference implies that models with one utilization variable to describe the productive process can generate nonsensical economic results if this variable is interpreted as representing either duration or speed, because the behaviour of costs can represent only one of the two interpretations. To illustrate, a recent body of literature relates capital utilization, economic growth and the speed of convergence (for example, Chatterjee 2005), by assuming depreciation to increase with utilization at an increasing rate. This makes sense if one

justifies increases in utilization as a result of increases in speed. Yet this literature justifies increases in utilization as a result of increases in duration through increases in the average workweek of capital.

Another interesting feature of the 'speed' model stems from the first-order conditions for cost minimization, which can be used to show that, if v, K and L are treated as choice variables, at the optimum, $r(v) = r'(v)v$. When duration and speed are endogenous this characteristic generalizes to $r(v, d) = r_v(v, d)\ v$ and optimal speed is determined by optimal duration (Madan 1987). This is consistent with the finding by Bresnahan and Ramey (1994) for the auto industry that line speed and the number of shifts are long-run margins of adjustment.

Consider now the representation of the productive process underlying the typical definitions of capacity utilization. Namely,

$$x = F(K, L) \tag{4}$$

where all variables are defined as before and speed and duration set at unity. Using (4), Panzar's (1976) definition of capacity becomes:

$$h(K) = \max_L F(K, L) \tag{5}$$

where $h(K)$ is an increasing function of K. This definition leads to an output-based definition of short-run capacity utilization; that is:

$$CU = x/x\ \max \tag{6}$$

where x max is given by (5).

When capital equipment is capacity-rated in terms of output units, as in electricity generation, one can measure directly the denominator of (6) and short-run capital and capacity utilization coincide (cf. Winston 1982, ch. 5). In general, however, the denominator in (6) is not well defined. An alternative procedure is to define the denominator in (6) as the optimal level of output, x^0. For instance, in the literature on dynamic factor demand models x^0 is defined as the optimal level of output when the capital stock is endogenous (for example, Morrison 1985; also see Prucha and Nadiri 1996, for a generalization). Since 'optimal' output varies with the specification of the optimization problem, one can generate a variety of reasonable definitions of capacity utilization which measure different concepts. Not surprisingly, the corresponding empirical definitions fail to move together (de Leeuw 1979) or with the average workweek of capital (Beaulieu and Mattey 1998).

Implications

Perhaps the most important economic implication of the analysis of capital utilization above is for our understanding of technical change at the aggregate level. Ignoring increases in duration understates the contribution of capital services to output growth and, thus, overstates the estimates of technical change or the Solow residual in standard sources of growth analysis. Beaulieu and Mattey's estimate of the annual rate of growth in the average workweek of capital for manufacturing over the 1974–91 period is 0.17. They use employment per shift as weights, which are the appropriate ones, and find that only 25 per cent of the variation in growth can be accounted for by overtime.

Macroeconomists have pursued this issue but emphasized its business cycle implications. That is, when the Solow residual is adjusted for the workweek of capital it ceases to be pro-cyclical. For instance, Shapiro (1993) made this point in a widely cited paper. His results continued to hold in Beaulieu and Mattey's more recent data and they have given rise to a substantial literature that we will not explore here. One implication of this finding noted by Shapiro is that it casts doubts on alternative explanations of the behaviour of the residual stressing market power when there are substantial costs to adjusting the workweek of capital, for example through the shift differential.

There is an early literature on the human costs of shift-work which may be captured through the shift differential. Betancourt and Clague (1981, ch. 12) conclude from their review of this literature that observed shift differentials of four to five

per cent in the United States substantially underestimate the human costs of shift-work. This conclusion is consistent with estimates in an unpublished paper by Shapiro (1995) that the marginal shift premium is 25 per cent. A strand of literature in labour economics on compensating differentials has considered shift-work.

Kostiuk (1990) obtains estimates of the shift differential of well above ten per cent in the unionized sector for both 1979 and 1985. He relies on Census of Population Survey data for his analysis.

An issue neglected in the recent literature is the role of obsolescence in capital utilization. Marris (1964) argued that an increase in the rate of obsolescence should strengthen the economic incentive for shift-work, since it ameliorated disincentive effects of wear and tear depreciation. In the last few decades we have observed systematic shifts from mechanical technologies to electronic technologies, which diminish wear and tear costs and increase the rate of obsolescence. This shift should, thus, have provided an incentive for increased capital utilization. Yet, to my knowledge, the economic literature has not addressed this issue explicitly.

Finally, an important reason for interest in capital utilization as an economic variable is the existence of transaction costs and market imperfections. These frictions make ownership of capital equipment and structures attractive relative to rentals for instantaneous capital services. Of course these rental markets do not exist in most cases. A substantial recent literature in industrial organization investigates the effect of transaction costs, including incompleteness of contracts and agency costs, on incentives and the evolution of institutions. With one exception, it has not addressed the impact of changes in transaction costs and market imperfections on capital utilization. The exception is the work of Hubbard (2003) on the trucking industry. He shows that improvements in monitoring technology in the form of on board computers increase capacity utilization, which in this industry coincides with short-run capital utilization just as in the electricity generation industry. Issues of long-run capital utilization and relevance for other industries, however, remain unexplored in this context.

See Also

- ▶ Adjustment Costs
- ▶ Fixed Factors
- ▶ Labour Market Institutions

C

Bibliography

Anxo, D., G. Bosch, D. Bosworth, G. Cette, T. Sterner, and D. Taddei. 1995. *Work patterns and capital utilization: An international comparative study*. London: Kluwer Academic Publications.

Beaulieu, J., and J. Mattey. 1998. The workweek of capital and capital utilization in manufacturing. *Journal of Productivity Analysis* 10: 199–203.

Betancourt, R. 1986. A generalization of modern production theory. *Applied Economics* 18: 915–928.

Betancourt, R., and C. Clague. 1981. *Capital utilization: A theoretical and empirical analysis*. New York: Cambridge University Press.

Betancourt, R., C. Clague, and A. Panagariya. 1985. Capital utilization and factor specificity. *Review of Economic Studies* 52: 311–329.

Bresnahan, T., and V. Ramey. 1994. Output fluctuations at the plant level. *Quarterly Journal of Economics* 108: 593–624.

Cardellichio, P. 1990. Estimation of production behavior using pooled micro data. *The Review of Economic and Statistics* 72: 11–18.

Chatterjee, S. 2005. Capital utilization, economic growth and convergence. *Journal of Economic Dynamics and Control* 29: 2093–2124.

Coates, D. 1991. Endogenous capital utilization and taxation of corporate capital. *National Tax Journal* 44: 79–91.

Das, S. 1992. A micro-econometric model of capital utilization and retirement: The case of the US cement industry. *Review of Economic Studies* 59: 277–297.

de Leeuw, F. 1979. Why capacity utilization rates differ. In *Measures of Capacity Utilization: Problems and Tasks*, ed. F. de Leeuw, et al. Staff Studies No. 105. Washington, DC: Board of Governors of the Federal Reserve System.

Foss, M. 1981. Long-run changes in the workweek of fixed capital. *American Economic Review, Papers and Proceedings* 71: 58–63.

Garcia Sanchez, A., and M. Vazquez Mendez. 2005. The timing of work in a general equilibrium model with shiftwork. *Investigaciones Economicas* 29: 149–179.

Georgescu-Roegen, N. 1972. Process analysis and the neoclassical theory of the firm. *American Journal of Agricultural Economics* 54: 279–294.

Hubbard, T. 2003. Information, decisions, and productivity: On-board computers and capacity utilization in trucking. *American Economic Review* 93: 1328–1353.

Kostiuk, P. 1990. Compensating differentials for shiftwork. *Journal of Political Economy* 98: 1054–1075.
Madan, D. 1987. Optimal duration and speed in the long run. *Review of Economic Studies* 54: 695–700.
Marris, R. 1964. *The economics of capital utilization*. Cambridge: Cambridge University Press.
Marshall, A. 1873.The future of the working classes. In *Memorials of Alfred Marshall*, ed. A.C. Pigou. London: Macmillan, 1925.
Marshall, A. 1923. *Industry and trade,* 4th ed. Reprints of economic classics, New York: Augustus M. Kelley, 1970.
Marx, K. 1867. *Capital*, vol. 1, New York: Vintage Books; Random House, 1979.
Mayshar, J., and Y. Halevy. 1997. Shiftwork. *Journal of Labor Economics* 15: S198–S222.
Morishima, M. 1973. *Marx's economics*. Cambridge: Cambridge University Press.
Morrison, C. 1985. On the economic interpretation and measurement of optimal capacity utilization with anticipatory expectations. *Review of Economic Studies* 52: 295–310.
Panzar, J. 1976. A neoclassical approach to peak load pricing. *The Bell Journal of Economics* 7: 521–530.
Prucha, I., and M. Nadiri. 1996. Endogenous capital utilization and productivity measurement in dynamic factor demand models. *Journal of Econometrics* 71: 343–379.
Shapiro, M. 1993. Cyclical productivity and the workweek of capital. *American Economic Review* 83: 229–233.
Shapiro, M. 1995. *Capital utilization and the marginal premium for work at night*. Mimeo: University of Michigan.
Smith, K. 1970. Risk and the optimal utilization of capital. *Review of Economic Studies* 37: 253–259.
Winston, G. 1982. *The timing of economic activity.* Cambridge: Cambridge University Press.

Capital, Credit and Money Markets

Benjamin M. Friedman

The markets for money, credit and capital represent a fundamental dimension of economic activity, in that the many and varied functions of the modern economy's financial markets both reflect and help shape the course of the economic system at large. Financial markets facilitate such central economic actions as producing and trading, earning and spending, saving and investing, accumulating and retiring, transferring and bequeathing. Development of the financial system is a recognized hallmark of economic development in the broadest sense.

Neither the important role played by the financial side of economic activity nor economists' awareness of it is a recent phenomenon. Economic analysis of the roles of money, credit and capital constitutes a tradition as old as the discipline itself. Nevertheless, in comparison with other equally central objects of economic analysis this tradition is as remarkable for its continuing diversity as for the richness of the insights it has generated. A century after Marshall and Wicksell and Bagehot, a half-century after Keynes and Robertson and Hicks, and a quarter-century after the initial path-breaking work of Tobin and Modigliani and Milton Friedman, there is still no firm consensus on many of the more compelling questions in the field: What are the most important determinants of an economy's overall level of capital intensity? How does risk affect the allocation of that capital? Do leverage and intermediation of debt matter for aggregate economic outcomes? Does money matter – and, if so, what is it?

The absence of universally accepted answers to these and other fundamental questions does not signify a failure to develop conceptual understanding of how the markets for money, credit and capital function, or of the basic elements of these markets' interactions with non-financial economic activity. The persistent diversity of thought on these unresolved questions has instead reflected the inability of empirical analysis, hindered by the continual and at times rapid evolution of actual financial systems, to provide persuasive evidence on issues characterized both by a multiplicity of plausibly relevant determining factors and by the inherent unobservability of some of the most important among them – for example, ex ante perceptions of risks as well as rewards.

The Market for Capital

The essential reason for having a capital market in any economy stems from the nature of the productive process. In all economies anyone has ever observed, and the more so in the more developed

among them, production of goods and services to satisfy human wants relies on capital as well as labour. If capital is to exist to use in production, someone must own it; and in economies in which this ownership function lies with individuals or other private entities, the primary initial role of the capital market is to establish the terms on which capital is held. In market-oriented economies the terms on which capital is (or may be) held provide incentives affecting the further accumulation of new capital, so that over time the capital market plays an additional, logically consequent role in determining the economy's existing amount of capital and hence its potential ability to produce goods and services.

In conceptualizing how the market mechanism sets the terms on which an economy's capital is held, economists have traditionally paired the role of capital as an input to the production process with the role of capital as a vehicle for conveying wealth – that is, ultimate command over goods and services – forward in time. The capital market is therefore the economic meeting place between the theory of production, often in the derivative form of the theory of investment, and the theory of consumption and saving. Different assumptions forming the underlying theory on either side in general lead to differing characterizations of how the capital market establishes the terms on which capital is held, and consequently differing characterizations of how the market affects the economy's accumulation of capital over time and hence its capital intensity at any point in time. Among the critical features of production theory and consumption-saving theory that have featured prominently in this analysis of their intersection are the substitutability of capital for other production inputs, the source and nature of technological progress, and the interest elasticity of saving. In most modern treatments, these specifics in turn depend on more basic assumptions like the respective specifications of the production function constraining producers and the intertemporal utility function maximized by wealth-holders.

Notwithstanding the central importance of this basic economic role of the capital market, as well as the insight and ingenuity with which economists over many years have elaborated their understanding of it, what gives the modern study of capital markets much of its particular richness is the focus on one particular factor that could, in principle, be entirely absent from this economic setting, but that is ever present in reality: uncertainty.

The essential feature of capital from this perspective is its durability. Because capital is durable – that is, its use in production does not instantly consume or destroy it – it provides those who hold it with not just the ability but the necessity to convey purchasing power forward in time in a specific form. Precisely because of this durability, capital necessarily exposes those who hold it to whatever uncertainties characterize both the production process and the demand for wealth-holding in the future.

Not just reward but risk too, therefore, are inherent features of capital that must accrue to some holders, somewhere in the economy, if the economy is to enjoy the advantages of production based in part on durable capital inputs. The introduction of risk has profound implications for consumption-saving behaviour. In addition, when the absence of perfect rental markets leads producers who use capital to be also among the holders of capital, the introduction of risk in this way affects production-investment behaviour too. Hence via at least one side of the capital market nexus, and via both sides under plausibly realistic assumptions, the risk consequent upon the durability of capital alters the determination of the terms on which capital is held, and thereby alters the determination of the economy's capital accumulation. Increasingly in recent years, the study of capital markets by economists has focused on the market pricing of this risk. The context in which this risk pricing of function matters, however, remains the consequences, for wealth-holding and for investment and production, of the terms on which capital is held.

The implications of the risk inherent in durable capital depend, of course, on many aspects of the capital market environment. Two prominent features of existing capital markets in particular have importantly shaped the explosive development of the capital markets risk-pricing literature during the past quarter-century. First, durable capital is

not the only available form of wealth holding. Other assets may be risky too, but at least some assets exist which do not expose holders to the risks, involving unknown outcomes far in the future, that are consequent on the durability of typical capital assets. Second, even capital assets are not all identical. Heterogenous capital assets expose their holders to risks that not only are not identical but also, in general, are not independent.

Following Markowitz (1952) and Tobin (1958), the investigation of the allocation of wealth-holding between a single risk-free asset and a single risky asset readily establishes the terms on which (risky) capital is held, in the form of the excess of its expected return over the known return on the alternative (presumed risk-free) asset. In the simplest case of a single-period-at-a-time decision horizon, for example, the maximization of utility exhibiting constant relative risk aversion in the sense of Pratt (1964) and Arrow (1965), subject to the assumption that the uncertain return to capital is normally distributed, leads to the result that an investor's demand for capital, expressed in proportion to the investor's total wealth, depends linearly on the expected excess return:

$$\frac{1}{w} \cdot A_K^D = \frac{1}{\rho \cdot \sigma_K^2} \cdot [E(r_K) - \overline{r}] \qquad (1)$$

where W is the investor's total wealth, A_K^D is the quantity demanded of the risky asset, ρ is the coefficient of relative risk aversion, $E(r_K)$ and σ_K^2 are respectively the mean and variance of the ex ante distribution describing assessments of the uncertain asset return, and $\overline{r}$ is the known return on the alternative asset. (This simple result is both convenient and standard, but it can be only an approximation because normally distributed asset returns are strictly incompatible with utility functions exhibiting constant relative risk aversion.) If it is possible to represent the economy's aggregate asset demands in a form corresponding to Eq. 1 for individual investors, then the requirement that the existing amount of each asset must equal to the amount demanded leads to the result that the expected excess return on capital depends linearly on the composition of the existing wealth:

$$E(r_K) = \overline{r} + \rho\sigma_K^2 \cdot \frac{A_K}{W} \qquad (2)$$

where A_K is the actual existing quantity of the risky asset. If the market equilibration process works via changes in the price of the risky asset, rather than its stated per-unit return, then both A_K and W are jointly determined with $E(r_K)$ and the resulting relationship is analogous though no longer linear:

$$E(r_K) = \overline{r} + \rho\sigma_K^2 \cdot \frac{P[E(r_K)] \cdot \overline{A}_K}{A_F + P[E(r_K)] \cdot \overline{A}_K} \qquad (3)$$

where A_F is the existing quantity of the risk-free asset (taken to have unit price), $\bar{A}_K$ is the quantity of the risky asset in physical units, and P is the price of the risky asset with $[\mathrm{d}P/\mathrm{d}E(r_K)] < 0$. (If capital is infinitely lived, $P = 1/E(r_K)$.) The addition of this element of the theory of risk pricing thus allows the capital market, in the context of a general economic equilibrium, to establish the terms on which durable capital is held – and hence the incentive to capital accumulation – when other, non-durable assets are also present.

The second major aspect of actual capital assets motivating the development of the economic analysis of capital markets is heterogeneity. Capital assets differ from one another not only because of actual physical differences but also because, with imperfect rental markets, the application of identical capital items to different uses in production has some permanence, so that ownership of a particular capital asset typically implies ongoing participation in a specific production activity. In general, each kind of capital asset, categorized not only by physical characteristics but also by production application, exposes those who hold it to a unique set of uncertainties. Moreover, in general the different risks associated in this way with different capital assets are not independent.

The elaboration of the single-risky-asset model in Eqs. 1, 2, and 3 due to Sharpe (1964) and Lintner (1965) readily represents the determination of relative returns in the capital market, in this

context of heterogeneous capital assets with interdependent risks, and hence enables the outcomes determined in the capital market to affect not just the aggregate quantity but also the allocation of the company's capital accumulation. The multivariate analogues of Eqs. 1 and 2 are simply

$$\frac{1}{W}\cdot A_K^D = \frac{1}{\rho}\Omega^{-1}[E(r_k) - \overline{r}\cdot 1] \tag{4}$$

$$E(r_K) = \overline{r}\cdot 1 + \rho\Omega\cdot\frac{1}{W}\cdot A_K \tag{5}$$

where A_K^D, $\boldsymbol{A}_K$ and $\boldsymbol{r}_K$ are vectors with individual elements respectively corresponding to A_K^D, $\boldsymbol{A}_K$ and r_K, $\boldsymbol{\Omega}$ is the variance-covariance structure associated with expectations $E(\mathbf{r}_K)$, and $\mathbf{1}$ is a vector of units. In Eq. 4 the demand for each specific capital asset depends linearly on the expected excess return over the risk-free rate not only of that asset but of all other capital assets as well, with the substitutability between any two assets – that is, the response of the demand for one asset to the expected return on another – determined by the investor's risk aversion as well as by the interdependence among the respective returns on all of the risky assets. In Eq. 5 the equilibrium expected excess return on each capital asset at any time therefore depends (linearly) on the existing quantities of all assets expressed as shares of the economy's total wealth. Under conventional models of investment behaviour, the accumulation of each specific kind of capital over time depends in turn on the entire set of equilibrium returns determined in this way.

Moreover, this role of the capital market in guiding the allocation of capital does not depend in any fundamental way on the presence of an alternative asset with risk-free return. If all assets bear uncertain returns, either because capital assets are the only existing assets, or because even the returns on other assets are uncertain (because of uncertain price inflation, for example), the analogue of Eq. 4 is

$$\frac{1}{W}\cdot A_K^D = \frac{1}{\rho}\left[\Omega^{-1} - \left(1'\Omega^{-1}1\right)^{-1}\Omega^{-1}1'\Omega^{-1}\right]\cdot E(r_K) + \left(1'\Omega^{-1}1\right)^{-1}\Omega^{-1}1. \tag{6}$$

The second term in Eq. 6 represents the composition of the minimum-variance portfolio, which in the absence of a risk-free asset is a unique combination of risky assets, expressed as a vector of asset shares adding to unity. The first term in Eq. 6 expresses the investor's willingness to hold a portfolio different from this minimum-variance combination. The transformation of $\boldsymbol{\Omega}$ contained in the first term maps what is in general a variance-covariance matrix of full rank into a matrix of rank reduced by one, as is implied by the balance sheet constraint emphasized by Brainard and Tobin (1968). Because the resulting matrix is of less than full rank, however, no exact analog of Eq. 5 then exists.

Combining the description of asset demands in Eq. 6 with the requirement of market clearing therefore determines the relative expected returns among all assets – in other words, determines the absolute expected returns on all assets but one, given the expected return on that one – but cannot determine absolute expected returns without at least some reference point fixed outside the risk pricing mechanism. This result is in fact analogous to the implication of Eq. 5 (or Eq. 3), in that Eq. 5 determines the expected return on each risky asset only in relation to the fixed benchmark of the known return on the alternative risk-free asset. In either case the analysis of risk pricing alone is insufficient to determine absolute returns without something else, presumably grounded in the fundamental interrelation between the respective roles of capital in production and in wealth-holding, to anchor the overall return structure.

Actual capital markets perform these functions of pricing risk and thereby guiding the accumulation and allocation of new capital, in essentially all advanced economies with well developed financial systems. In most such economies, the most immediately visible focus of the risk pricing mechanism is the trading on stock exchange of existing claims to capital in the form of equity ownership shares in ongoing business enterprises. Equity shares are composite capital assets not only in the sense that each business firm typically owns a variety of different kinds of physical capital but also because the value of most firms consists in part of intangible capital in the form of

existing knowledge, organization and reputation. In the context of what are often very large costs of establishing new enterprises, together with highly imperfect secondary markets for physical capital assets, even in principle the prices of equity securities need not correspond in any direct way to the liquidation value of a firm's separate items of plant and equipment. Given transactions costs and imperfect secondary markets, the existing enterprise itself is just as much an aspect of an advanced economy's long-lived production technology as is the sheer physical durability of capital.

Markets in which existing equity shares are traded also present the opportunity for the initial sale to investors of new equity shares issued by business enterprises in order to augment their available financial resources. In addition to guiding capital accumulation and allocation by establishing the relevant risk pricing, therefore, capital markets also play a direct role in facilitating capital accumulation by offering firms the opportunity to raise new equity funds directly. Even so, given firms' ability to increase their equity base by retaining their earnings rather than distributing them fully to shareholders – and also given the availability of debt financing (see the discussion of credit markets immediately below) – the extent to which firms actually rely on new issues of equity varies widely from one economy to another. In the United States, for example, well established firms typically do not issue new equity shares in significant volume, and the market for new issues is primarily a resource for new enterprises of a more speculative character. (The aggregate net addition to equity in the US market each year is typically negative, in that equity retirements and repurchases exceed gross new issues.) In most other economies, too, new issues of equity shares provide only small amounts of net funds for business.

Even when new equity additions via new shares issues are small, however, the risk pricing function of the capital market still guides an economy's capital accumulation and allocation process. Internal additions to equity from retained earnings are by far the major source of equity funds for the typical business in most economies, and – at least in theory – the retention or distribution of earnings by firms reflects in part considerations of expected return and associated risk as priced in the capital markets. Firms in lines of business in which new investment is less profitable (after allowance for risk) than the economy's norm not only cannot issue new equity shares on attractive terms but also must either distribute their earnings or face undervaluation of their outstanding shares by market investors. Conversely, firms with unusually profitable prospects at the margin of new investment can favourably issue new shares or can retain their earnings to fund their expansion.

Finally, two further features of actual modern capital markets bear explicit notice. Each, appropriately considered, is consistent with the notion of capital markets serving the basic function of pricing risk, and thereby guiding an economy's capital accumulation and allocation.

First, highly developed capital markets are characterized by enormous volumes of trading. In principle, the risk-pricing mechanism could function with little trading of existing securities, and under the right conditions it could function with none at all. If investors all agreed on the appropriate set of price relationships, there would be neither the incentive nor the need to effect actual transactions. The agreed-upon set of prices might fluctuate widely or narrowly, depending upon changes in assessments of risk and return, but as long as the assessments were universally shared there would be little if any trading.

The huge trading volumes typical of actual modern capital markets therefore suggest that, in fact, investors do not share identical risk and return assessments. Annual trading volume on the New York Stock Exchange, for example, is normally near one-half the total value of listed existing shares. Although the continually changing circumstances of both individual and institutional investors no doubt play some role, it is difficult to explain this phenomenon except in the context of substantial heterogeneity in the response of investors' risk and return assessments to the flow of new information.

The possibility that investors' opinions differ is only a minor complication for the theory of risk

pricing as sketched above. Lintner (1969) showed that competitive capital markets with heterogeneous investors determine outcomes for the pricing of risky assets that just reflect an appropriately constructed aggregation over all individual investors' differing assessments (as well as their differing preferences), weighted by their respective wealth positions. The question remains, however, why investors' assessments differ. One line of analysis, initiated by Grossman (1976), has emphasized systematic differences in assessments due to underlying differences in information available to different investors. By contrast, Shiller (1984) suggested the importance of unsystematic differences not readily explainable within the conventional analytic framework based on rational maximization. The question remains unsettled but important nonetheless.

The second additional feature of actual modern capital markets that bears explicit attention is the proliferation of increasingly complex securities, including options, warrants, futures, and so forth. Given heterogeneity among investors, this development fits naturally in the context of the capital markets' basic economic role of establishing the terms on which the risks inherent in a capital-intensive production technology are to be borne. When investors differ among themselves in age, or wealth, or preferences, or risk and return assessments, in general the most efficient allocation of those risks does not consist of all investors' holding portfolios embodying identical risks and prospective returns. Instead, different investors will hold differing portfolios, and a further role of an economy's capital markets is to allocate the bearing of specific risks across different investors.

Heterogeneity among different kinds of physical assets would itself facilitate such specialization, and heterogeneity among the business enterprises whose equity shares constitute the asset units in actual capital markets typically does so to an even greater extent. Still, even this resulting degree of feasible specialization in risk bearing apparently falls well short of what would be fully consistent with the existing extent of investor heterogeneity.

Complex securities enable the capital markets to achieve a more efficient allocation of risk across heterogeneous investors by more finely dividing the risk inherent in an economy's production technology. Options, for example, permit an investor not merely to hold a (positive or negative) position in the equity of a specific firm but to hold positions corresponding only to designated parts of the distribution describing the possible outcomes for that firm's performance as reflected in the price of its equity shares. While the existing array of complex securities presumably does not approach the set of contingent claims necessary to span the space of possible outcomes in the sense of Arrow (1964) and Debreu (1959), developments along these lines in recent years have presumably rendered risk bearing more efficient. Moreover, following Merton (1973a) and Black and Scholes (1973), the analysis of the market pricing of risk has extended to explicitly contingent claims the central features of market equilibrium. The analysis is richer, therefore, and the outcome more efficient, but the end result of the economic process remains the pricing of the risk associated at any time with the existing stock of capital, with consequent effects on the total accumulation and allocation of capital over time.

The Market for Credit

The presence of heterogeneity among different participants in a market economy also provides an economic rationale for credit markets. The primary initial role of the credit market is to facilitate borrowing and lending – that is, the transfer of purchasing power by the issuing and acquiring (and trading) of money-denominated debts. In establishing the terms on which such transfers take place, the credit market plays a role in guiding the allocation of the economy's resources that is parallel to that played by the capital market.

If all market participants were identical, such a market could establish terms on which the representative agents would be willing to borrow or lend, but no actual borrowing or lending would take place. Under those circumstances the credit market would be of little economic importance. By contrast, actual economies consist of an almost infinite variety of differently positioned

participants. Individuals differ from business enterprises, and private-sector entities differ from governments. Even just among individuals, there are old and young, rich and poor, highly and weakly risk-averse, favourably and unfavourably taxed, home-owners and renters, and so on in ever more dimensions and ever greater detail. As a result, the credit market does not just establish a putative price for strictly hypothetical trades. It facilitates transfers that in turn make possible resource allocations which could not otherwise come about.

At the most basic level, economists since Fisher (1930) have emphasized the role of borrowing and lending in achieving a separation between production and consumption decisions. Here the function of the credit market is to enable individuals to shift purchasing power forward or backward in time, so as to free the timing pattern of consumption streams from the corresponding timing pattern of earnings from production (while still preserving, of course, the relevant constraint connecting the appropriately discounted totals). The overall result of this intertemporal separation is, in general, to achieve more efficient resource allocations in the sense both of greater production from given available inputs as well as higher utility from given available consumption. Without such a separation it would be impossible to construe the intertemporal theory of consumption and saving as in any way distinct from the theory of production and investment. Even the limited heterogeneity between firms and households is sufficient to give rise to borrowing and lending along these lines.

Nevertheless, the question of why money-dominated debts should serve this intertemporal transfer function – rather than having all obligations take the form of direct ownership claims to capital, for example – opens up a whole series of further important issues. Following the analysis of capital markets immediately above, the most readily apparent answer is that debt obligations isolate the specific risks associated with the purchasing power of the unit of denomination (in other words, inflation risk) and risks associated with the borrower's ability to meet the stated obligation (default risk), and that this conventional compartmentalization is evidently convenient for a variety of reasons. Inflation risk and default risk are in general not independent, however. In addition, it is just as easy to imagine alternative conventions that might be just as convenient, like the predominant use of debts denominated in purchasing-power units.

Given the conventional monetary denomination of debt obligations, the function of the credit market in most modern economies is to redistribute immediate claims to purchasing power, in exchange for future claims, along three major dimensions of heterogeneity: between individuals and firms, between the private sector and the government, and between domestic and foreign entities. In addition, redistributions among individuals (and, to a lesser extent, among firms) are often a further important credit market function.

Business firms typically apply to investment not only their equity additions from retained earnings and any new share issues but also funds raised by borrowing. Modigliani and Miller (1958) set forth conditions under which the firm's reliance on debt versus equity financing would be a matter of indifference, in that it would not affect the firm's total value, but conditions prevailing in actual economies and their capital and credit markets do not meet these conditions closely. Business reliance on debt financing is typically large, and it varies systematically across countries and across industries within a given country. Prominent aspects of the divergence of actual economies from the Modigliani-Miller irrelevance conditions which the ensuing voluminous literature has emphasized, include tax structures, risks and costs of bankruptcy by the firm, differential borrowing rates for firms and individuals (due to, for example, risks and costs of bankruptcy by individuals), monitoring costs required to minimize risks, and restrictive features of debt contracts intended to reduce risks due to moral-hazard effects of imperfectly compatible incentive structures.

The resulting substantial reliance on debt financing by business means that credit markets, like capital markets, play a major economic role in guiding an economy's accumulation and allocation of capital over time. When any or all of the

factors cited above lead business enterprises to finance a new investment with some combination of additional equity (from retained earnings or new share issues) and additional debt, the appropriate calculation of investment incentives involves the cost to the firm in both the capital market and the credit market. In circumstances in which the financing margin corresponding to marginal new investment is a debt margin – as is often the case in the United States, for example, where firms' reliance on external funds is typically synonymous with issuance of debt – the relevant cost at the margin is the cost in the credit market.

Use of the credit market to finance government spending is among the oldest and most prevalent forms of financial transactions, and it has, understandably, generated an entire literature unto itself. In practical terms, government reliance on the credit markets in most modern economies is important not only in that governments often issue debt to finance large portions of their total spending but also because government borrowing often absorbs a large mount of the total funds advanced in the market by lenders. As is the case for private borrowers, government debt issues separate in time the ability to spend from the need to raise revenue. In addition, however, because under some circumstances governments need not repay debt obligations at all (they may refinance them forward indefinitely), and also because of uncertainty over the identity of the responsible taxpayers even in the case of future repayment, government debt is in part net wealth to the aggregate of private holders in a way that private debts are not.

The distinguishing feature of government debt in many economies is its essential freedom from default risk. In addition, in most economies the market for government debt is among the most efficiently functioning of all financial markets. Hence the existence of government debt enables the credit market to establish a base, with risk factors limited to inflation and real discounting values, from which it can then price privately issued debts subject to risks associated with default as well. The practice of giving government guarantees to the payment of interest and principal on selected private debts, which has greatly proliferated in recent years, has further increased the variety of forms of default-free debt securities. Yet another important implication of the default-free nature of government debt is that, to the extent that government borrowing takes the place of borrowing that individuals could do on their own account only at higher cost or not at all, government debt is in part net wealth to the private sector even if it is necessarily repaid and even if the identity of the responsible taxpayers is fully known.

International borrowing and lending has also greatly increased in recent years, as technological advances in communications have brought the world's financial markets closer together in the relevant physical sense, while individual countries' governments have progressively relaxed legal and regulatory barriers that impede international capital flows. From the perspective of any one country, the possibility of international borrowing and lending serves a separation function analogous to the fundamental Fisherian separation of production and consumption decisions in a closed economy. An economy that can borrow or lend abroad need not balance its imports and exports at each moment of time. Moreover, once an economy builds up a positive net international creditor position, it can indefinitely finance an excess of imports over exports from the associated interest income. (Conversely, once an economy builds up a net international debtor position, it must indefinitely export in excess of its imports so as to finance the debt service.) From the perspective of the world economy as a whole, international borrowing and lending is even more closely analogous to the closed economy model, in that it facilitates a more efficient allocation of resources across national boundaries.

Apart from these categorical heterogeneities, credit markets also reallocate immediate purchasing power among individuals and among business firms. The need for individuals in differing circumstances to make a complementary arrangements for divergences among their respective income and spending streams is basic to any life-cycle or overlapping-generations model of consumer behaviour. On the borrowing side, practical market limitations on individuals' issuance of

equity-type claims contingent on their future earnings means that the only effective way for most individuals to shift command over purchasing power from the future to the present is through ordinary money-denominated debts. In fact, in most economies individuals' ability to borrow against no security other than future earnings is severely limited in any form, so that most borrowing by individuals occurs in conjunction with the purchase of homes, automobiles or other specific durable goods. On the lending side, individuals choosing to carry purchasing power into the future can hold wealth in any of its available forms, and in fact most individuals hold by far the greater part of their wealth in forms other than credit market instruments. Hence the great bulk of the borrowing done by individuals represents funds advanced by financial intermediary institutions rather than directly by other individuals.

Direct borrowing and lending among business firms is also a significant part of credit market activity especially in highly developed financial systems. On the borrowing side, firms' reliance on debt finance is readily understandable for reasons sketched above, irrespective of whether the funds raised come from individuals, from financial intermediaries or from other businesses. On the lending side, debt held by business firms usually takes the form of very shortterm liquid instruments intended to provide maximum flexibility in the future disposition of the purchasing power thus deferred.

In sum, the credit markets play the fundamental role of enabling an economy populated by heterogeneous agents to achieve superior resource allocations by redistributing immediate purchasing power in exchange for money-denominated claims on the future. Because of the intensive use of debt to finance both business and residential investment, in establishing the terms on which such transfers take place also play a consequent role in guiding the economy's capital accumulation and capital allocation over time that is analogous to – and, in some economies, as important as – the parallel incentives provided by the capital markets. In addition, in part because those elements of total spending that are typically debt-financed bulk large in aggregate demand, in many economies fluctuations of overall economic activity are as closely related to the movement of total credit as to the movements of any other financial aggregates (like any measure of money, for example).

Finally, as in the case for capital markets, several other features of actual credit markets that in principle need not be so, but in fact are so, have exerted a strong influence on the way in which economists have studied these markets over many years. One of the most important in this regard is the fact, noted above, that individuals directly hold relatively few credit market instruments. Instead, the great bulk of the borrowing and lending in any even moderately advanced economy takes place through specialized financial intermediaries, including commercial banks, non-bank thrift institutions, insurance companies, pension funds, mutual funds, and so on.

Standard rationales underlying financial intermediation include the minimization of information and transactions costs, and the diversification of risks, in a world in which assets are imperfectly divisible and both asset returns and wealth-holders' cash-flow positions are imperfectly correlated. In principle, these rationales apply to capital markets as well as credit markets, and in many countries institutions like mutual funds and pension funds do play an important role in holding equity shares. In practice, however, in many countries the bulk of the existing equity securities is still held directly by individuals rather than through financial intermediaries, while the opposite is true for debt instruments. As a result, the study of financial intermediation in general, and of specific kinds of intermediary institutions in particular, has been a major focus of the economic analysis of credit markets.

Another feature of actual credit markets that has likewise attracted a voluminous economic literature has been the simultaneous existence of a great variety of different debt instruments, especially including debts that differ according to their respective stated maturities. Although in principle only a single form of debt instrument, with a unique maturity, would enable the credit market to serve much of its economic functions, in fact almost all known credit markets are characterized

by the simultaneous existence of many debt instruments with differing terms to maturity. The need for the market to price these debts – that is, to establish a term structure of interest rates – not only raises issues of risk analogous to those discussed above in relation to capital markets but also makes explicit the need for a more general intertemporal framework of analysis.

At least since Hicks (1939), economists have been aware at some level that short-term and long-term debts are both risky assets, each from a particular time perspective. Apart from risks associated with default and inflation, short-term debt provides a certain return to holders over a short-time horizon, so that short-term government debt could plausibly constitute the risk-free asset in a no-inflation version of the standard capital asset pricing model represented by Eqs. 1 and 2 above. Over a longer horizon, however, short-term debt preserves capital value only by exposing both borrowers and lenders to an income risk if interest rates fluctuate. Conversely, long-term debt maintains income streams only by exposing borrowers and lenders to the risk of fluctuating capital value over any time horizon shorter than the stated term to maturity. At an a priori level, there is no way to establish which form of risk is more important, and hence no way to establish even the sign of the expected return premium that risk-averse borrowers and lenders would establish in pricing short-term and long-term debts relative to one another.

Following both Hicks and Keynes (1936), most economists have assumed as an empirical matter that typically prevailing preferences are such that lenders require, and borrowers are willing to pay, a positive expected return premium for the capital risk inherent in long-term debt. Hence the subsequent development of the term structure literature has taken a form at least in principle compatible with the single-period capital asset pricing model. More recently, however, following Stiglitz's (1970) explicit demonstration of the connection between the risk pricing of receipt streams and preferences with respect to consumption streams, the economic literature of asset pricing has tended to return to the position that there is no general answer to the question of whether short-term or long-term debts are more risky. Instead, the preferred form of analysis has increasingly become an explicitly intertemporal model, like Merton's (1973b) intertemporal capital asset pricing model or, more recently, Ross's (1976) arbitrage pricing model as generalized by Cox et al. (1985).

Money Markets

The economic role played by the money market is more difficult to establish than that of the markets for capital and credit, in part because 'money' is not straightforward to define. The standard practice among non-economists, which often creates unexpected confusion for economists, is to refer to 'money' indistinguishably from short-term forms of credit, so that 'the money market' is just that segment of the credit market devoted to issuing and trading short-term debts, and 'money rates' are correspondingly the stated nominal interest rates on money market instruments thus defined. By contrast, economists have traditionally viewed, money as distinct from credit, and have given money a central place in macroeconomic analysis which typically appeals to some form of aggregation argument to assume away the existence of credit altogether.

Two lines of thinking, neither necessarily easy to convert into an operational definition of 'money', have traditionally dominated economists' thinking on the subject. One has emphasized the role of money as a form of wealth (in traditional language, a store of value). The problem then is to define which forms of wealth constitute money and which do not. The emphasis in drawing such distinctions has typically rested on the safety and liquidity of the asset, in the sense of its relative freedom from default risk and its ease of conversion, at a predetermined rate of exchange, into whatever is the economy's means of payment. Although the general idea behind such thinking is clear enough, in actually existing economies it has proved impossible to draw the requisite line between money and non-money assets without imposing arbitrary distinctions. Typically, the more highly developed an

economy's financial system, the greater is the need for such arbitrary judgements.

The alternative line of traditional thinking has been to emphasize the role of money in effecting transactions, and hence to define as money just those assets that are acceptable as means of payment. One problem here is that both legalities and common business practice sometimes make ambiguous what constitutes an acceptable means of payment. Indeed, in highly developed financial systems an increasing volume of transactions is effected without requiring the actual holding of any specific asset identifiable as money. Moreover, this approach leads to further difficulties, even apart from definitional problems. If money is used as one side of every transaction in the respective markets for all goods and services and all other assets, then the meaning of 'the money market' is unclear except in the sense that there exists a demand for money equal to the net supply of all other tradeables, and, correspondingly, a supply of money equal to the net demand for all other tradeables.

Under either the store-of-value approach or the means-of-payment approach, the central role conventionally attached to the money market in modern macroeconomic analysis primarily reflects the standard institutional structure within which monetary policy consists in the first instance of actions by the central bank that, either directly or through the financial intermediary system, affect the supply of money however it is defined. Market equilibrium then requires a corresponding change in the demand for money – that is, in the demand for highly liquid assets or for the means of payment, depending on the definitional approach assumed. In either case, the required shift in the public's aggregate portfolio demands presumably requires, in turn, a shift in the structure of expected asset returns, with consequent implications for non-financial economic activity under any of a variety of familiar theories of consumption, investment and production behaviour.

The specifics of this process, however, depend crucially on the definition of 'money'. Under the approach that identifies money with assets meeting sufficient criteria of safety and liquidity, the demand for money is merely a by-product of the theory of risk-averse portfolio selection under uncertainty. Under this approach, what is more difficult is to specify the process connecting the supply of money, so defined, to the central bank's actions. To the extent that the supply of assets defined as money consists largely of the liabilities of depository intermediaries, and to the extent that the relevant institutional arrangements require intermediaries to hold reserves against their liabilities, the connection between money supply and central bank actions that provide or withdraw intermediary reserves is apparent enough. When there is no reserve requirement, however – because either specific kinds of intermediary institutions or specific kinds of intermediary liabilities face no reserve requirement – the connection between monetary policy actions and money supply is more problematic.

The situation under the approach that identifies money with the means of payment is roughly the opposite. Because most economies' means of payment consist largely of the direct liabilities of the central bank and the reservable liabilities of specific intermediaries, connecting the supply of money to central bank reserve actions is relatively straightforward. What is more difficult under this approach is establishing the link to the demand for money thus defined, and hence ultimately the effect on non-financial economic activity. When assets other than the means of payment also provide safety and liquidity, the standard theory of portfolio selection no longer suffices to determine the demand for the means of payment itself. Economic analysis of this problem has largely developed along the inventory-theoretic lines laid out initially by Baumol (1952) and Tobin (1956) and by Miller and Orr (1966). Especially in modern circumstances that readily permit transactions on a credit basis, however, the relevance of such 'cash in advance' models is unclear.

Regardless of the specific conceptual approach taken to define money, it is clear that the deposit liabilities of financial intermediaries bulk large in individuals' direct wealth holding in most actual economies, so that economists' study of money markets has heavily focused on the role of

intermediaries and intermediation. The reasons for the prominent position of intermediary liabilities in individuals' direct wealth-holdings are not difficult to understand. The deposits of banks and similar intermediaries typically provide the most convenient means of settling most transactions, and the asset transformation provided by financial intermediations makes it attractive for most individuals to participate in the market for many kinds of assets via intermediaries rather than directly.

As a result, 'the money market' in most actual economies consists largely of financial intermediaries on one side and both individuals and business firms on the other. Here, as elsewhere in modern economies, the profusion of differentiated financial products is vast. Money market assets in this sense consist of checkable and non-checkable deposits, demand deposits and deposits for stated terms ranging from a few days to many months, deposits with fixed (nominal) returns and variable returns, and so on. Moreover, in the eyes of most market participants, short-term credit market claims that are close portfolio substitutes for intermediary deposits (commercial paper) are money market instruments too.

See Also

- ▶ Credit
- ▶ Finance
- ▶ Financial Intermediaries
- ▶ Financial Markets
- ▶ Monetary Policy
- ▶ Money Supply

References

Arrow, K.J. 1964. The role of securities in the optimal allocation of risk-bearing. *Review of Economic Studies* 31: 91–96.

Arrow, K.J. 1965. *Aspects of the theory of risk-bearing*. Helsinki: The Yrjo Jahnsson Foundation.

Baumol, W.J. 1952. The transactions demand for cash: An inventory theoretic approach. *Quarterly Journal of Economics* 66: 545–556.

Black, F., and M. Scholes. 1973. The pricing of options and corporate liabilities. *Journal of Political Economy* 81(3): 637–654.

Brainard, W.C., and J. Tobin. 1968. Pitfalls in financial model-building. *American Economic Review: Papers and Proceedings* 58: 99–122.

Cox, J.C., J.E. Ingersoll Jr., and S.A. Ross. 1985. A theory of the term structure of interest rates. *Econometrica* 53(2): 385–407.

Debreu, G. 1959. *Theory of value: An axiomatic analysis of economic equilibrium*. New Haven: Yale University Press.

Fisher, I. 1930. *The theory of interest*. New York: The Macmillan Company.

Grossman, S.J. 1976. On the efficiency of competitive stock markets when traders have diverse information. *Journal of Finance* 31(2): 573–585.

Hicks, J.R. 1939. *Value and capital*. Oxford: Oxford University Press.

Keynes, J.M. 1936. *The general theory of employment, interest and money*. London/New York: Macmillan/Harcourt, Brace & World.

Lintner, J. 1965. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *Review of Economics and Statistics* 47: 13–37.

Lintner, J. 1969. The aggregation of investors' diverse judgements and preferences in purely competitive securities markets. *Journal of Financial and Quantitative Analysis* 4(4): 347–400.

Markowitz, H. 1952. Portfolio selection. *Journal of Finance* 7: 77–91.

Merton, R.C. 1973a. Theory of rational option pricing. *Bell Journal of Economics and Management Science* 4(1): 141–183.

Merton, R.C. 1973b. An intertemporal capital asset pricing model. *Econometrica* 41(5): 867–887.

Miller, M.H., and D. Orr. 1966. A model of the demand for money by firms. *Quarterly Journal of Economics* 80: 413–435.

Modigliani, F., and M.H. Miller. 1958. The cost of capital, corporation finance, and the theory of investment. *American Economic Review* 48: 261–297.

Pratt, J.W. 1964. Risk aversion in the small and in the large. *Econometrica* 32: 122–136.

Ross, S.A. 1976. The arbitrage theory of capital asset pricing. *Journal of Economic Theory* 13(3): 341–360.

Sharpe, W.F. 1964. Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance* 19: 425–442.

Shiller, R.J. 1984. Stock prices and social dynamics. *Brookings Papers on Economic Activity* 2: 457–510.

Stiglitz, J.E. 1970. A consumption-oriented theory of the demand for financial assets and the term structure of interest rates. *Review of Economic Studies* 37(3): 321–351.

Tobin, J. 1956. The interest-elasticity of transactions demand for cash. *Review of Economics and Statistics* 38: 241–247.

Tobin, J. 1958. Liquidity preference as behavior toward risk. *Review of Economic Studies* 25: 65–86.

Capitalism

Robert L. Heilbroner

Abstract

Capitalism is a unique historical formation with core institutions and distinct movements. It involves the rise of a mercantile class, the separation of production from the state, and a mentality of rational calculation. Its characteristic logic revolving around the accumulation of capital reflects the omnipresence of competition. It displays broad tendencies to unprecedented wealth creation, skewed size distributions of enterprise, large public sectors, and cycles of activity. Whereas students of capitalism traditionally envisaged an end to the capitalist period of history, modern economists show little interest in historical projection.

Keywords

Baran, P.; Bentham, J.; Braudel, F.; Capital; Capital accumulation; Capitalism; Commodities; Comparative systems approach; Competition; Economic freedom; Employment contract; Engels, F.; English Revolution; Exchange value; Feudalism; French Revolution; Hirschman, A.; Ideology; Inequality; Keynes, J.; Labour; Labour power; Locke, J.; Marshall, T.; Marx, K.; Mercantilism; Mill, J. S.; Morality; Myrdal, G.; Polanyi, K.; Political economy; Political freedom; Pre-capitalist social formations; Private property; Profit; Property rights; Public expenditure; Public sector; Quasi-rent; Religion; Ricardo, D.; Role of the state; Schumpeter, J.; Smith, A.; Socialism; Surplus value; Use value; Veblen, T.; Weber, M.

JEL Classifications

P1

Capitalism is often called *market society* by economists, and the *free enterprise system* by business and government spokesmen. But these terms, which emphasize certain economic or political characteristics, do not suffice to describe either the complexity or the crucial identificatory elements of the system. Capitalism is better viewed as a historical 'formation', distinguishable from formations that have preceded it, or that today parallel it, both by a core of central institutions and by the motion these institutions impart to the whole. Although capitalism assumes a wide variety of appearances from period to period and place to place – one need only compare Dickensian England and 20th-century Sweden or Japan – these core institutions and distinctive movements are discoverable in all of them, and allow us to speak of capitalism as a historical entity, comparable to ancient imperial kingdoms or to the feudal system.

The most widely acknowledged achievement of capitalist societies is their capacity to amass wealth on an unprecedented scale, a capacity to which Marx and Engels paid unstinting tribute in *The Communist Manifesto*. It is important to understand, however, that the wealth amassed by capitalism differs in quality as well as quantity from that accumulated in precapitalist societies. Many ancient kingdoms, such as Egypt, displayed remarkable capacities to gather a surplus of production above that needed for the maintenance of the existing level of material life, applying the surplus to the creation of massive religious or public monuments, military works or luxury consumption. What is characteristic of these forms of wealth is that their desirable attributes lay in the specific use-values – war, worship, adornment – to which their physical embodiments directly gave rise. By way of decisive contrast, the wealth amassed under capitalism is valued not for its specific use-values but for its generalized exchange-value. Wealth under capitalism is therefore typically accumulated as *commodities* – objects produced for sale rather than for direct use or enjoyment by their owners; and the extraordinary success of capitalism in amassing wealth means that the production of commodities makes possible a far greater expansion of wealth than its accumulation as use-values for the rulers of earlier historical formations.

Both Smith and Marx stressed the importance of the expansion of the commodity form of

wealth. For example, Smith considered labour to be 'productive' only if it created goods whose sale could replenish and enlarge the national fund of capital, not when its product was intrinsically useful or meritorious. In the same fashion, Marx described the accumulation of wealth under capitalism as a circuit in which money capital (M) was exchanged for commodities (C), to be sold for a larger money sum (M′), in a never-ending metamorphosis of M–C–M′.

Although the dynamics of the M–C–M′ process vary greatly depending on whether the commodities are trading goods or labour power and fixed capital equipment, the presence of this imperious internal circuit of capital constitutes a prime identificatory element for capitalism as a historical genus. As such, it focuses attention on two important aspects of capitalism. One of these concerns the motives that impel capitalists on their insatiable pursuit. For modern economists the answer to this question lies in 'utility maximization', an answer that generally refers to the same presumed attribute of human nature as that which Smith called the 'desire of bettering our condition'. The unappeasable character of the expansive drive for capital suggests, however, that its roots lie not so much in these conscious motivations as in the gratification of unconscious drives, specifically the universal infantile need for affect and experience of frustrated aggression. Such needs and drives surface in all societies as the desires for prestige and for personal domination. From this point of view, capitalism appears not merely as an 'economic system' knit by the appeals of mutually advantageous exchange, but as a larger cultural setting in which the pursuit of wealth fulfils the same unconscious purposes as did the pursuit of military glory or the celebration of personal majesty in earlier epochs. Such a description conveys the force of the 'animal spirits' (as Keynes referred to them) that both set into motion, and are appeased by, the M–C–M′ circuit. (Heilbroner 1985, ch. 2; Sagan 1985, chs 5, 6).

A second general question raised by the centrality of the M–C–M′ circuit concerns the manner in which the process of capital accumulation organizes and disciplines the social activity that surrounds it. Here analysis focuses on the institutions necessary for the circuit to be maintained. The crucial capitalist institution is generally agreed to be private property in the means of production (not in personal chattels, which is found in all societies). The ability of private property to organize and discipline social activity does not however lie, as is often supposed, in the right of its owners to do with their property whatever they want. Such a dangerous social licence has never existed. It inheres, rather, in the right accorded its owners to withhold their property from the use of society if they so wish.

This negative form of power contrasts sharply with that of the privileged elites in precapitalist social formations. In these imperial kingdoms or feudal holdings, disciplinary power is exercised by the direct use or display of coercive force, so that the bailiff or the seneschal are the agencies through which economic order is directly obtained. The social power of capital is of a different kind – a power of refusal, not of assertion. The capitalist may deny others access to his resources, but he may not force them to work with them. Clearly, such power requires circumstances that make the withholding of access an act of critical consequence. These circumstances can only arise if the general populace is unable to secure a living unless it can gain access to privately owned resources or wealth. Capital thus becomes an instrument of power because its owners can establish claims on output as their *quid pro quo* for permitting access to their property.

Access to property is normally attained by the relationship of 'employment' under which a labourer enters into a contract with an owner of capital, usually selling a fixed number of working hours in exchange for a fixed wage payment. At the conclusion of this 'wage-labour' contract both parties are quit of further obligation to one another, *and the product of the contractual labour becomes* the property of the employer. From this product the employer will pay out his wage obligations and compensate his other suppliers, retaining as a profit any residual that remains.

In detail, forms of profit vary widely, and not all forms are specific to capitalism – trading gains,

for example, long predate its rise. Explanations of profit vary as a consequence, but as a general case it can be said that all profits depend ultimately on inequality of economic position. When the inequality arises from wide disparities of knowledge or access to alternative supplies, profits typically emerge as the mercantile gains that were so important in the eyes of medieval commentators, or as the depredations of monopolistic companies against which Adam Smith inveighed. When the inequality stems from differentials in the productivity of resources or productive capability we have the quasi-rents to which such otherwise different observers as Marshall and Schumpeter attribute the source of capitalist gain. And when the inequality is located in the market relationship between employer and worker it appears as the surplus value central to Marxian and, under a different vocabulary, to classical political economy. As Smith put it, 'Many a workman could not subsist a week, few could subsist a month, and scarce any a year without employment. In the long-run the workman may be as necessary to his master as his master is to him; but the need is not so immediate' (Smith [1776] 1976, p. 84).

This is not the place to enter into a discussion of these forms of profit, all which can be discerned in modern capitalist society. What is of the essence under capitalism is that gains from whatever origin are assigned to the owners of capital, not to workers, managers or government officials. This is a clear indication both of the difference of capitalism from, and its resemblance to earlier social formations. The difference is that product itself now flows to owners of property who have already remunerated its producers, not to its producers – usually peasants in precapitalist societies – who must then 'remunerate' their lords. The resemblance is that both arrangements channel a social surplus into the hands of a superior class, a fact that again reveals the nature of capitalism as a system of social domination, not merely of rational exchange.

Thus we can see that the successful completion of the circuit of accumulation represents a political as well as an economic challenge. The attainment of profit is necessary for the continuance of capitalism not alone because it replenishes the wherewithal of each individual capitalist (or firm) but because it also demonstrates the continuing validity and vitality of the principle of M–C–M′ as the basis on which the formation can be structured. Profit is for capitalism what victory is for a regime organized on military principles, or an increase in the number of adherents for one built on a proselytizing religion.

The Evolution of Capitalism

Capitalism as a 'regime' whose organizing principle is the ceaseless accumulation of capital cannot be understood without some appreciation of the historic changes that bring about its appearance. In this complicated narrative it is useful to distinguish three major themes. The first concerns the transfer of the organization and control of production from the imperial and aristocratic strata of precapitalist states into the hands of mercantile elements. This momentous change originates in the political rubble that followed the fall of the Roman empire. There merchant traders established trading niches that gradually became loci of strategic influence, so that a merchantdom very much at the mercy of feudal lords in the 9th and 10th centuries became by the 12th and 13th centuries an estate with a considerable measure of political influence and social status. The feudal lord continued to oversee the production of the peasantry on his manorial estate, but the merchant, and his descendant the guild master, were organizers of production in the towns, of trade between the towns and of finance for the feudal aristocracy itself.

The transformation of a merchant estate into a capitalist class capable of imagining itself as a political and not just an economic force required centuries to complete and was not, in fact legitimated until the English revolution of the 17th and the French revolution of the 18th centuries. The elements making for this revolutionary transformation can only be alluded to here in passing. A central factor was the gradual remonetization of medieval European life that accompanied its political reconstitution. The replacement of feudal social relationships, mediated through custom and

tradition, by market relationships knit by exchange worked steadily to improve the wealth and social importance of the merchant against the aristocrat. This enhancement was accelerated by many related developments – the inflationary consequence of the importation of Spanish gold in the 16th century, which further undermined the rentier position of feudal lords; the steady stream of runaway serfs who left the land for the precarious freedom of the towns and cities, placing further economic pressure on their former masters; the growth of national power that encouraged alliances between monarchs and merchants for their mutual advantage; and yet other social changes (see Pirenne 1936; Hilton 1978).

The overall transfer of power from aristocratic to bourgeois auspices is often subsumed under the theme of the rise of market society; that is, as the increasingly *economic* organization of production and distribution through purchase and sale rather than by command or tradition. This economic revolution, from which emerge the 'factors of production' that characterize market society, must however be understood as the end product of a *political* convulsion in which one social order is destroyed to make way for a new one. Thus the creation of a propertyless waged labour force – the prerequisite for the appearance of labour-power as a commodity that would become enmeshed in the M–C–M′ circuit – is a disruptive social change that begins in England in the late 16th century with the dispossession of peasant occupants from communal land and does not run its course until well into the 19th century. In similar fashion, the transformation of feudal manors from centres of social and juridical life into real estate, or the destruction of the protected guilds before the unconstrained expansion of nascent capitalist enterprises, embody wrenching socio-political dislocations, not merely the smooth diffusion of preexisting economic relations throughout society. It is such painful rearrangements of power and status that underlay the 'great transformation' out of which capitalist market relationships finally arise (Polanyi 1957, Part II).

A second theme in the historical evolution of capital emphasizes a related but distinct aspect of political change. Here the main emphasis lies not so much in the functional organization of production as in the separation of a traditionally seamless web of rulership, extending over all activities within the historical formation, into two realms, each concerned with a differentiated part of the whole. One of these realms involved the exercise of the traditional political tasks of rulership – mainly the formation and enforcement of law and the declaration and conduct of war. These undertakings continued to be entrusted to the existing state apparatus which retained (or regained) the monopoly of legal violence and remained the centre of authority and ceremony. The other realm was limited to the production and distribution of goods and services; that is, to the direction of the material affairs of society, from the marshalling of the workforce to the amassing and use of the social surplus. In the fulfilment of this task, the second realm also extended its reach beyond the boundaries of the territorial state, insofar as commodities were sold to and procured from outlying regions and countries that became enmeshed in the circuit of capital.

The formation of these two realms was of epoch-making importance for the constitution of capitalism. The creation of a broad sphere of social activity from which the exercise of traditional command was excluded bestowed on capitalism another unmistakable badge of historic specificity; namely, the creation of an 'economy', a semi-independent state within a state and also extending beyond its borders.

This in turn brought two remarkable consequences. One of these was the establishment of a political agenda unique to capitalism, in which the relationship of the two realms became a central question around which political discussion revolved, and indeed continues to revolve. In this discussion the overarching unity and mutual dependency of the two realms tends to be overlooked. The organization of production is generally regarded as a wholly 'economic' activity, ignoring the political function performed by the wage–labour relationship in disciplining the workforce in lieu of bailiffs and seneschals. In like fashion, the discharge of political authority is regarded as essentially separable from the operation of the economic realm, ignoring the provision

of the legal, military and material contributions without which the private sphere could not function properly or even exist. In this way, the presence of two realms, each responsible for part of the activities necessary for the maintenance of the social formation, not only gives to capitalism a structure entirely different from that of any pre-capitalist society but also establishes the basis for a problem that uniquely preoccupies capitalism; namely, the appropriate role of the state vis-à-vis the sphere of production and distribution.

More widely recognized is the second major effect of the division of realms in encouraging economic and political freedom. Here the capitalist institution of private property again takes centre stage, this time not as a means of arranging production or allocating surplus, but as the shield behind which designated personal rights can be protected. Originally conceived as a means for securing the accumulations of merchants from the seizures of kings, the rights of property were generalized through the market into a general protection accorded to all property, including not least the right of the worker to the ownership of his or her own labour-power.

Now the wage–labour relationship appears not as means for the subordination of labour but for its emancipation, for the crucial advance of wage-labour over enslaved or enserfed labour lies in the right of the working person to deny the capitalist access to labour-power on exactly the same legal basis as that which enables the capitalist to deny the worker access to property. There is, therefore, an institutional basis for the claim that the two realms of capitalism are conducive to certain important kinds of freedom, and that a sphere of market ties may be necessary for the prevention of excessive state power. This is surely an important part of Smith's celebration of the society of 'natural liberty', and has been the basis of the general conservative endorsement of capitalism. Unquestionably, the greatest achievements of human liberty thus far attained in organized society have been achieved in certain advanced capitalist societies. One cannot, however, make the wider claim that capitalism is a sufficient condition for freedom, as the most cursory survey of modern history will confirm.

A third theme in the evolution of capitalism calls attention to the cultural changes that have accompanied and shaped its institutional framework. Much emphasis has been given to this theme in the work of Weber and Schumpeter, both of whom stress the historic distinctions between the essentially rational – that is, means-ends calculating – culture of capitalist civilization compared with the 'irrational' cultures of previous social formations. Here it is important to recognize that rationality does not refer to the *principle* of capitalism, for we have seen that the impetus to amass wealth is only a sublimation of deeper-lying non-rational drives and needs, but to the behavioural paths followed in the pursuit of that principle. The drive to amass capital can be analysed in terms of a calculus that is less readily apparent, if indeed present at all, in the search for other forms of prestige and power. This pervasive calculating mind-set is itself the outcome both of the abstract nature of exchange-value, which makes possible commensurations that cannot be carried out in terms of glory or sheer display, and of the pressures exerted by the marketplace, which penalize economic actors who fail to follow the arrow of economic advantage. Capitalism is therefore distinguishable in history by the predominance of a prudent, accountant-like comparison of costs and benefits, a perspective discoverable in the mercantile pockets of earlier formations but highly uncharacteristic of the tempers of their ruling elites (see Weber 1930; Schumpeter 1942, ch. XI).

The cultural change associated with capitalism goes further, however, than the rationalization of its general outlook. Indeed, when we examine the general culture of capitalist life we are most forcibly struck by an aspect that precedes and underlies that highlighted above. This is the presence of an ideological framework that contrasts sharply with that of pre-capitalist formations. I do not use the word *ideology* in a pejorative sense, as denoting a set of ideas foisted on the populace by a ruling order in order to manipulate it, but rather as a set of belief systems to which the ruling elements of the society themselves turn for self-clarification and explanation. In this sense, ideology expresses what the dominant class in a society

sincerely believes to be the true explanations of the questions it faces.

That which is characteristic of the ideologies of earlier formations is their unified and monolithic character. In the ancient civilizations of which we know, an all-embracing world view, usually religious in nature, explicates every aspect of life, from the workings of the physical universe, through the justification of rulership, down to the smallest details of social routines and attitudes. By way of contrast, the ideology that emerges within capitalism is made up of diverse strands, more of them secular than religious and many of them in some degree of conflict with other strands. By the end of the 18th century, and to some degree before, the explanation system to which capitalist societies turn with respect to the workings of the universe is science, not religious cosmology. In the same manner, rulership is no longer regarded as the natural prerogative of a divinely chosen elite but perceived as 'government'; that is, as the manner in which 'individuals' create an organization for their mutual protection and advancement. Not least, the panorama of work and the patterns of material life are perceived not as the natural order of things but as a complex web of interactions that can be made comprehensible through the teachings of political economy, later economics. The individual threads of these separate scientific, political-individualist and economic belief systems originate in many cases before the unmistakable emergence of capitalism in the 18th century, but their incorporation into a skein of culture provides yet another identifying theme of the history of capitalist development.

Within this skein, the ideology of economics is obviously of central interest for economists. A crucial element of this belief system involves changes in the attitude towards acquisitiveness itself, above all the disappearance of the ancient concern with good and evil as the most immediate and inescapable consequence of wealth-gathering. As Hirschman has shown, this change was accomplished in part by the gradual reinterpretation of the dangerous 'passion' of avarice as a benign 'interest', capable of steadying and domesticating social intercourse rather than disrupting and demoralizing it (Hirschman 1977). Other crucial elements of understanding were provided by Locke's brilliant demonstration in *The Second Treatise on Government* (1690) that unlimited acquisition did not contravene the dictates of reason or Scripture, and by the full pardon granted to wealth-seeking by Bentham, who demonstrated that the happiness of all was the natural outcome of the self-regarding pursuit of the happiness of each.

The problem of good and evil was thus removed from the concerns of political economy and relegated to those of morality; and economics as an inquiry into the workings of daily life was thereby differentiated from earlier inquiries, such as the reflections of Aristotle or Aquinas, by its explicit disregard of their central search for moral understanding. Perhaps more accurately, the constitution of a 'science' of economics as the most important form of social self-scrutiny of capitalist societies could not be attempted until moral issues, which defied the calculus of the market, were effectively excluded from the field of its investigations.

The Logic of the System

This conception of capitalism as a historical formation with distinctive political and cultural as well as economic properties derives from the work of those relatively few economists interested in capitalism as a 'stage' of social evolution. In addition to the seminal work of Marx and the literature that his work has inspired, the conception draws on the writings of Smith, Mill, Veblen, Schumpeter and a number of sociologists and historians, notable among them Weber and Braudel. The majority of present-day economists do not use so broad a canvas, concentrating on capitalism as a market system, with the consequence of emphasizing its functional rather than its institutional or constitutive aspects.

In addition to the characteristic features of its institutional 'nature', capitalism can also be identified by its changing configurations and profiles as it moves through time. Insofar as these movements are rooted in the behaviour-shaping properties of its nature, we can speak of them as

expressing the logic of the system, much as conquest or dynastic alliance express the logic of systems built on the principle of imperial rule, or the relatively changeless selfvreproduction of primitive societies expresses the logic of societies ordered on the basis on kinship, reciprocity and adaptation to the givens of the physical environment.

The logic of capitalism ultimately derives from the pressure exerted by the expansive M–C–M′ process, but it is useful to divide this overall force into two categories. The first of these concerns the 'internal' changes impressed upon the formation by virtue of its necessity to accumulate capital – its metabolic processes, so to speak. The second deals with its larger 'external' motions – changes in its institutional structure or in important indicia of performance as the system evolves through history.

The internal dynamics of capitalism spring from the continuous exposure of individual capitals to capture by other capitalists. This is the consequence of the disbursement of capital-as-money into the hands of the public in the form of wages and other costs. Each capitalist must then seek to win back his expended capital by selling commodities to the public, against the efforts of other capitalists to do the same. This process of the enforced dissolution and uncertain recapture of money capital in the circuit of accumulation is, of course, the pressure of competition that is the social outcome of generalized profit-seeking. We can see, however, that competition cannot be adequately described merely as the vying of suppliers in the marketplace. As both Marx and Schumpeter recognized, competition is at bottom a consequence of the mutual encroachments bred by the capitalist drive for expansion, not of the numbers of firms contending in a given market.

The process of the inescapable dissolution and problematical recapture of individual capitals now gives rise to the activities designed to protect these capitals from seizure. The most readily available means of self-defence is the search for new processes or products that will yield a competitive advantage – the same search that also serves to facilitate the expansion of capital through the development of new markets. Competition thus reinforces the introduction of technological and organizational change into the heart of the accumulation process, usually in two forms: attempts to cheapen the cost of production by displacements of labour by machinery (or of one form of fixed capital by another); or attempts to gain the public's purchasing power by the design of wholly new forms of commodities. As a consequence, one of the most recognizable attributes of capitalist 'internal' dynamics has been its constant revolutionizing of the techniques of production and its continuous commodification of material life, the sources of its vaunted capacity to change and elevate living standards.

A further internal change also arises from the expansive pressures of the core process of capital accumulation. This is a threat to the capacity as a whole to extract a profit from the production of commodities. This tendency arises from the long-run effect of rising living standards in strengthening the bargaining power of labour versus capital. There is no way in which individual enterprises can ward off this threat by cutting wages, for in a competitive market system they would thereupon lose their ability to marshall a workforce. Their only protection against a rising tendency of the wage level is to substitute capital for labour where that is possible. For the system as a whole, the need to hold down the bargaining power of labour must therefore hinge on a generalization of individual cost-reducing efforts, through the system-wide displacement of labour by machinery, or by the direct use of government policies to maintain a profit-yielding balance between labour and capital, or by systemic failures – 'crises' – that create generalized unemployment. Whether attempted by deliberate policy or brought about by the outcome of spontaneous market forces, the pressure to secure a profit-compatible level of wages thus becomes a key aspect in the internal dynamics of the system.

A final attribute of the internal logic of capitalism must also be traced to its core process of accumulation. This is the achievement of a highly adaptive method of matching supplies against demands without the necessity of political intervention. This cybernetic capacity is

surely one of the historical hallmarks of capitalism, and is regularly emphasized in the 'comparative systems approach' in which the responsive capacities of the market mechanism are compared with the inertias and rigidities of systems in which tradition or command (planning) must fulfil the allocational task. A critique of the successes and failures of the market system cannot be attempted here. Let us only emphasize that the workings of the system itself derive from institutional attributes whose genesis we have already observed – namely, the establishment of free contractual relations as the means for social coordination; the establishment of a social realm of production and distribution from which government intervention is largely excluded; the legitimation of acquisitive behaviour as the social norm; and activating the whole, the imperious search for the enlargement of exchange-value as the active principle of the historical formation itself.

Large-Scale Tendencies

From the metabolism of capitalism also emerges its larger 'external' motions – the overall trajectory often described as its macroeconomic movement, and the configurational changes that are the main concern of institutional economics. It may be possible to convey some sense of these general movements if we note three general aspects characteristic of them.

We have already paid heed to the first of these, the tendency of the capitalist system to accumulate wealth on an unparalleled scale. Some indication of the magnitude of this process emerges in the contrast between the increase in per capital GNP of developed (capitalist) and less-developed (noncapitalist) countries (Table 1):

Capitalism, Table 1 GNP per capita (1960 dollars and prices)

	Presently developed countries	Presently less-developed countries
Around 1750	$180	$180–90
Around 1930	780	190
Around 1980	3000	410

Source: Paul Bairoch in Faaland (1982), p. 162

After our lengthy discussion of the central role of accumulation within capitalism it does not seem necessary to relate this historic trend to its institutional base. Two somewhat neglected aspects of the overall increase in wealth seem worth mentioning, however. The first is that the increase in per capita GNP includes both augmentations in the volume of output and an extension of the M–C–M′ process itself within the social world. This is manifested in a continuous implosion of the accumulation process within capitalist societies – the process of the commodification of material life to which we earlier referred – and its explosion into neighbouring noncapitalist societies.

This explosive thrust calls attention to the second attribute of the overall expansion of wealth. It is that capital, as such, knows no national limits. From its earliest historic appearance, capital has been driven to link its 'domestic' base with foreign regions or countries, using the latter as suppliers of cheap labour-power or cheap raw materials or as markets for the output of the domestic economy. The consequence has been the emergence of self-reinforcing and cumulative tendencies towards strength at the centre, to which surplus is siphoned, and weakness in the periphery, from which it is extracted. The economic dimensions of this global drift are immediately visible in the previous table. This is the basis for what has been called the 'development of underdevelopment' as the manner in which ancient patterns of international hegemony are expressed in the context of capitalist relationships (Myrdal 1957, Part I: Baran 1957, chs V–VII).

We turn next to a different overall manifestation of the larger logic of capitalist development – its changes in institutional texture. There have been, of course, many such changes in the long span of Western capitalist experience – indeed, it is the very diversity of the faces of capitalism that prompted our search for its deep-lying identifying elements. Nonetheless, two changes deserve to be singled out, not only because of their sweeping magnitude and transnational occurrence, but because they have deeply altered the evolutionary

logic of the system itself. These have been the emergence within all modern capitalisms of highly skewed size distributions of enterprise, and of very large and powerful public sectors.

The general extent of these transformations is sufficiently well known not to require detailed exposition here. Suffice it to illustrate the trend by contrasting the largely atomistic composition of manufacturing enterprise in the United States at the middle of the 19th century with the situation in the 1980s, when seven-eighths of all industrial sales were produced by 0.1 per cent of the population of industrial firms. The enlargement of the public sector is not so dramatic but is equally unmistakable. During the present century in the United States, its size (measured by all government purchases of output plus transfer payments) has increased from perhaps 7.5 per cent of GNP to over 35 per cent, a trend that is considerably outpaced by a number of European capitalisms.

The first of these two large-scale shifts in the configuration can be directly traced to the pressures generated by the M–C–M′ circuit. The change from a relatively homogeneous texture of enterprise to one of extreme disparities of size is the consequence not only of differential rates of growth of different units of capital, but of defensive business strategies of trustification and merger, and the winnowing effect of economic disruptions on smaller and weaker units of capital. There is little disagreement as to the endemic source of this transformation in the dynamics of the marketplace and the imperative of business expansion.

The growth of large public sectors is not so immediately attributable to the accumulation process proper but rather results from changes in the logic of capitalist movements after the concentration of industry has taken place. Here the crucial change lies in the increasing instability of the market mechanism, as its constituent parts cease to resemble a honeycomb of small units, individually weak but collectively resilient, and take on the character of a structure of beams and girders, each very strong but collectively rigid and interlocked. It seems plausible that this rigidification was the underlying cause of the increasingly disruptive nature of the crises that appeared first in the late 19th century and climaxed in the great depression of the 1930s; and it is widely accepted that the growth of the public sector mainly owes its origins to efforts to mitigate the effects of that instability or to prevent its recurrence.

This brings us to the last general aspect of capitalist development; namely, the tendency for interruptions and failures to break the general momentum of capital accumulation. Perhaps no aspect of the logic of capitalism has been more intensively studied than these recurrent failures in the accumulation process. In the name of stagnation, gluts, panics, cycles, crises and long waves a vast literature has emerged to explain the causes and effects of intermittent systematic difficulties in successfully negotiating the passage from M to M'. The variables chosen to play strategic roles in the explanation of the phenomenon are also widely diverse: the saturation of markets; the undertow of insufficient consumption; the technological displacement of labour; the pressure of wages against profit margins; various monetary disorders; the general 'anarchy' of production; the effect of ill-considered government policy, and still others.

Despite the variety of elements to which various theorists have turned, a common thread unites most of their investigations. This is the premise that the instabilities of capitalist growth originate in the process of accumulation itself. Even theorists who have the greatest confidence in the inherent tendency of the system to seek a steady growth path, or who look to government intervention (in modern capitalism) as the main instability-generating force, recognize that economic expansion tends to generate fluctuations in the rate of growth, whether from the 'lumpy' character of investment, volatile expectations, or other causes. In similar fashion, economists who stress instability rather than stability as the intrinsic tendency of the system do not deny the possibility of renewed accumulation once the decline has performed its surgical work; indeed, Marx, the most powerful proponent of the inherently unstable character of the M–C–M′ process, was the first to assert that the function of crisis was to prepare the way for a renewal of accumulation.

In a sense, then, the point at issue is not whether economic growth is inherently unstable,

but the speed and efficacy of the unaided market mechanism in correcting its instability. This ongoing debate mainly takes the form of sharp disagreements with respect to the effects of government policy in supplementing or undermining the corrective powers of the market. The failure to reach accord on this issue reflects more than differences of informed opinion with regard to the consequences of sticky wages or prices, or ill-timed government interventions, and the like. It should not be forgotten that, from the viewpoint of capitalism as a regime, interruptions pose the same threats as did hiatuses in dynastic succession or breakdowns of imperial hegemony in earlier formations. It is not surprising, then, that the philosophic predilections of theorists play a significant role in their diagnoses of the problem, inclining economists to one side or the other of the debate on the basis of their general political sympathies with the regime, rather than on the basis of purely analytic considerations.

Periodization and Prospects

All the foregoing aspects of the system can be traced to its inner metabolism, the money–commodity–money circuit. This is much less the case when we now consider the overarching pattern of change described by the configuration of the social formation as a whole as it moves from one historic 'period' to another.

Traditionally these periods have been identified as early and late mercantilism; pre-industrial, and early and late industrial capitalism; and modern (or late, or state) capitalism. These designations can be made more specific by adumbrating the kinds of institutional change that separate one period from another. These include the size and character of firms (trading companies, putting-out establishments, manufactories, industrial enterprises of increasing complexity); methods of engaging and supervising labour (cottage industry through mass production); the appearance and consolidation of labour unions within various sectors of the economy; technological progress (tools, machines, concatenations of equipment, scientific apparatus); organizational evolution (proprietorships, family corporations, managerial bureaucracies, state participation). David Gordon has coined the term 'social structure of accumulation' to call attention to the changing framework of technical, organizational and ideological conditions within which the accumulation process must take place. Gordon's concept, applied to the general problem of periodization, emphasizes the manner in which the accumulation process first exploits the possibilities of a 'stage' of capitalism, only to confront in time the limitations of that stage which must be transcended by more or less radical institutional alterations (Gordon 1980).

The idea of an accumulation process alternately stimulated and blocked by its institutional constraints provides an illumining heuristic on the intraperiod dynamics of the system, but not a theory of its long-run evolutionary path. This is because not all national capitalisms make the transitions with equal ease or speed from one social structure to another, and because it is not apparent that the pressures of the M–C–M′ process push the overall structure in any clearly defined direction. Thus Holland at the end of the 17th century failed to make the leap beyond mercantilism, and England in turn in the second half of the 19th century failed to create a successful late industrial capitalism. In this regard it is interesting that the explanatory narratives of the great economists apply with far greater cogency to the evolutionary trends within periods than across them – Smith's scenario of growth in *The Wealth of Nations,* for instance, containing no suggestion that the system would move into an industrial phase with quite different dynamics, or Marx's depiction of the laws of motion of the industrialized system containing no hint of its worldwide evolution towards a state-underwritten structure. Although the inner characteristics of the M–C–M′ process enable us to apply the same generic designation of capitalism to its successive species-forms, it does not seem to be possible to demonstrate, even after the fact, that the transition from one stage to another had to be made, or to predict before the fact what the direction of institutional adjustment will be.

These cautions apply to the prospectus confronting capitalism in our day. Its long post World War II boom seems to have been based on three attributes of the social structure of accumulation of that time. One of these was the increasing interconnection between the political and the economic realms, not merely to provide a public base for mass consumption but to utilize the state's power of finance and international leadership to promote foreign private trade and production. Japanese capitalism has been the much cited case in point for the latter development. A second characteristic of the boom was the extraordinary development of technology, based on the close integration of scientific research and technical application. A third was the pronounced bourgeoisification of working-class life, especially in Europe and Japan, greatly reducing the spectre of class conflict in capitalist politics.

On the basis of these developments capitalism enjoyed the longest uninterrupted period of accumulation in its history, from the early 1950s to the mid-1970s. Not only was the boom uninterrupted save for minor and shortlived recessions, but on the wings of its new technological breakthroughs, and under the auspices of its active state cooperation, capitalism made extraordinary advances in introducing its core institutions into many areas of the underdeveloped world.

This halcyon period came to a sharp end in 1980 when growth rates in the United States and Europe fell precipitously. Some, although not all of the causes of this depression can be ascribed to an exhaustion of the expansionary possibilities within the postwar social structure of accumulation. The effect of enlarged and sustained public expenditure gradually shifted from the encouragement of production to the inducement of inflation, thus setting the stage for the adoption of the tight money policies that finally broke the back of the boom. As markets became saturated, the advances in technology lost their capacity to stimulate capital expansion and attention was increasingly directed to their system-threatening aspects – ecologically dangerous products, employment-eroding processes and sovereignty- defying enhancements of the international mobility of money capital and commodities. The international character of capital acquired extraordinary importance, as multinational corporations transplanted fixed capital into underdeveloped regions, from which it launched artillery barrages of commodities back on its domestic territory. And not least, the bourgeoisification of labour may have removed a traditional source of adaptational pressure from capitalism.

It is not possible to foretell how these challenges will be met, or what institutional changes will be forced upon the capitalist world as their consequence, or which capitalist nations will find the institutional and organizational means best suited to continue the accumulation process in this newly emerging milieu. Thus there is no basis for predicting the longevity of the social formation, either in its national instantiations or as a formational whole.

But while history forces on us a salutary agnosticism with regard to the longterm prospects for capitalism, it is interesting to note that all the great economists have envisaged an eventual end to the capitalist period of history. Smith describes the accumulation process as ultimately reaching a plateau when the attainment of riches will be 'complete', followed by a lengthy and deep decline. Ricardo and Mill anticipate the arrival of a 'stationary state', which Mill foresees as the staging ground for a kind of associationist socialism. Marx anticipates a series of worsening crisis, each crises serving a temporary rejuvenating function but bringing closer the day when the system will no longer be able to manage its internal contradictions. Keynes foresees ' a somewhat comprehensive socialization of investment'; Schumpeter, an evolution into a kind of bureaucratic socialism. By way of contrast, contemporary mainstream economists are largely uninterested in questions of historic projection, regarding capitalism as a system whose formal properties can be modelled, whether along general equilibrium or more dynamic lines, without any need to attribute to these models the properties that would enable them to be perceived as historic regimes and without pronouncements as to the likely structural or political destinations towards which they incline. At a time when the need for institutional adaptation seems pressing, such an

historical indifference to the fate of capitalism, on the part of those who are professionally charged with its self-clarification, does not augur well for the future.

See Also

► Socialism

Bibliography

Baran, P. 1957. *The political economy of growth*. New York: Monthly Review Press.
Faaland, J. 1982. *Population and the world economy in the 21st century.* Oxford: Blackwell.
Gordon, D. 1980. Stages of accumulation and long economic cycles. In *Processes of the world system*, ed. T. Hopkins and I. Wallerstein. Beverly Hills: Sage.
Heilbroner, R.L. 1985. *The nature and logic of capitalism*. New York: W.W. Norton.
Hilton, R., ed. 1978. *The transition from feudalism to capitalism*. London: Verso.
Hirschman, A. 1977. *The passions and the interests*. Princeton: Princeton University Press.
Myrdal, G. 1957. *Rich lands and poor.* New York: Harper & Bros.
Pirenne, H. 1936. *Economic and social history of Medieval Europe*. London: K. Paul, Trench, Trubner & Co..
Polanyi, K. 1957. *The great transformation*. Boston: Beacon Press.
Sagan, E. 1985. *At the dawn of tyranny: The origins of individualism, political oppression, and the state*. New York: Knopf.
Schumpeter, J. 1942. *Capitalism, socialism and democracy.* New York: Harper & Bros.
Smith, A. 1776. *The wealth of nations*. Oxford: Clarendon Press. 1976.
Weber, M. 1930. *The protestant ethic and the spirit of capitalism*. London: Allen & Unwin.

Capitalistic and a Capitalistic Production

Lionello F. Punzo

If 'capital' is the set of produced means of production, (almost) all production is capitalistic. Thus, the presence of capital in this sense can at most be (and in the history of economic doctrines was taken to represent) a necessary condition for defining capitalistic production. Differences arose as to the relative emphasis put on the social or techno-economic aspects of such transformation processes.

In Marx's analysis, capitalistic production is the organization of social production specific to a society characterized by private ownership of the means of production and by its separation from 'labour'. This historically given Mode of Production is contrasted with pre- and post-capitalistic forms, where power relationships are regulated according to different principles. By contrast, the distinction between production with and without capital focuses upon the relationship between means and objectives (consumption goods) of production activity. It played a role in the era of the full articulation of neoclassical thought. Its analytical use obviously depended upon the specific conception (and representation) of capital.

According to perhaps the most common theory, Capital is a factor of production, a member of a triad with Labour and Land. This view emphasizes the aspect of capital as a stock of man-produced goods which are at any point of time available in fixed quantities. (A)capitalistic production entails the application of (un)aided labour to natural resources. On the other hand, according to the Austrian (Böhm-Bawerk) definition, capital is the set of intermediate goods (or 'maturing consumption goods') emerging in the transformation of labour services into final goods when indirect methods of production are employed. This conception emphasizes the functional relationship whereby capital is the mode of realization of advanced production activity. Accordingly, acapitalistic production is direct production of consumption goods through application of bare labour to natural resources. Finally, in Wicksell's theoretical compromise, capital is a stock of used-up services of both labour and land. Production without capital is carried on by means of labour and natural resources in a state where capital goods either do not exist or are free goods relative to the available technology. (See Part II of the first volume of Wicksell's Lectures,

1934.) This definition obviously overlooks the fact that capital goods are themselves a byproduct of the advancement of technological knowledge, an idea implicit in Böhm-Bawerk and hinted at by Schumpeter.

At any rate, in all its various interpretations, acapitalistic production was a logical abstraction meant to illustrate, in a simpler analytical context, some basic principles holding for capitalistic production. In Böhm-Bawerk, this is the principle of the higher productivity of indirect (i.e. capitalistic) methods of production. In Wicksell, the distinction is meant to illustrate the marginalistic approach to the distribution of income and to show how it can be extended from the simpler production with labour and land only to production involving capital goods. In the former case, wage rate and rent are regulated by the marginal productivities of the two factors, in a state of full employment of labour and zero entrepreneurial profits. However, the extension of the marginal productivity principle to the theory of interest meets a crucial conceptual difficulty due to the fact that capital, being an aggregate of produced goods, has to be measured in value and the latter depends itself on income distribution. It is to avoid a circular argument that Wicksell proposes to regard capital as 'a single coherent mass of saved up resources'. Hence, interest would be (equal to) the difference between the marginal productivity of saved up labour and land and the marginal productivity of current labour and land. According to Wicksell, 'experience' shows that capital has a higher productivity and this is the reason why its share in the national product is normally positive.

It has been proved, in the debate on capital theory in the 1960s, that Wicksell's attempt at finding a way out of the difficulties of the marginalistic approach to income distribution is unsatisfactory. However, the recurrence of the theme of the distinction between acapitalistic and capitalistic production is interesting for it indicates the neoclassical authors' awareness of the theoretical difficulties they met in the treatment of capital and distribution.

See Also

- ▶ Capital Perversity
- ▶ Wicksell, Johan Gustav Knut (1851–1926)

Bibliography

Schumpeter, J.A. 1954. *History of economic analysis*. London: Allen & Unwin.

von Böhm-Bawerk, E. 1889. *Positive Theorie des Kapitals*. Trans. G.D. Huncke, vol. 2, *Capital and interest*. South Holland: Libertarian Press, 1959.

Wicksell, K. 1934. In *Lectures on political economy*, 1st English ed, ed. L. Robbins. London: George Routledge & Sons.

Carey, Henry Charles (1793–1879)

Henry W. Spiegel

American social scientist. Born in Philadelphia, the son of Mathew Carey, he was a prolific author, and his influence, though short-lived, spread from Pennsylvania throughout the nation and to Europe.

Carey's economic views were sharply at variance with those of Ricardo and Malthus, and reflect the optimism characteristic of American conditions favourable to economic expansion, conditions from which Carey himself benefited as a successful entrepreneur and promoter. The two leading themes of his writings were protectionism and harmony of interests. In his first book, *Essay on the Rate of Wages* (1835), he opposed trade restrictions as running counter to the providential order. But in *The Past*, *the Present*, *and the Future* (1848) and in later writings, he vigorously appealed for tariff protection as fulfilling his law of association, a law that called for diversified and balanced regional development. Narrow specialization and foreign trade would violate this law. In *The Slave Trade* (1853) Carey suggested protectionism for the South, where it would foster industrial development.

The scope of Carey's optimistic belief in a harmonious order gradually widened. In his first book he postulated harmony between capitalists and workers, the former benefiting from rising profits and the latter from wages that rose as a result of the accumulation of capital. In his *Principles of Political Economy* (1837–1840) the landowner becomes part of the harmonious order, with his earnings depicted as a return on his capital rather than a gift of nature. Population growth does not disturb the harmony as it is restrained by social conditioning. There are further attacks against the Ricardian rent theory in *The Past*, *the Present*, *and the Future*, where cultivation is said to move from inferior to superior land, not vice versa as Ricardo had taught, and with returns increasing rather than decreasing. In the *Principles of Social Science* (1858–1859) Carey expands his vision of a harmonious order to apply to the universe, and in *The Unity of Law* (1872) he maintains that cosmic and social laws are identical. Carey has been characterized as 'easily the most perverse and the most original American political economist before Veblen' (Conkin 1980, p. 261).

Selected Works

1835. *Essay on the rate of wages*. Philadelphia: Carey, Lea & Blanchard.
1837–1840. *Principles of political economy*, 3 vols. Pennsylvania: Carey, Lea & Blanchard.
1838. *The credit systems in France, Great Britain and the United States*. Philadelphia: Carey, Lea & Blanchard.
1848. *The past, the present, and the future*. Philadelphia: Carey & Hart; London: Longman, Brown, Green, and Longmans.
1851. *The harmony of interests, agricultural, manufacturing and commercial*. Philadelphia: J.S. Skinner; 2nd ed., New York: M. Finch, 1852.
1853. *The slave trade, domestic and foreign: Why it exists, and how it may be extinguished*. Philadelphia: A. Hart.
1858–1859. *Principles of social science*, 3 vols. Philadelphia: J.B. Lippincott & Co.; London: Trübner & Co.
1863. *Financial crises: Their causes and effects*. Philadelphia: Baird.
1867. *Reconstruction: Industrial, financial & political*. Philadelphia: Collins.
1872. *The unity of law; as exhibited in the relations of physical, social, mental and moral science*. Philadelphia: H.C. Baird.

References

Conkin, P.K. 1980. *Prophets of prosperity: America's first political economists*. Bloomington: Indiana University Press.
Dorfman, J. 1946. *The economic mind in American civilization 1606–1865*, vol. 2. New York: Viking.
Green, A.W. 1951. *Henry Charles Carey: Nineteenth-century sociologist*. Philadelphia: University of Philadelphia Press.
Kaplan, A.D.H. 1931. *Henry Charles Carey: A study in American economic thought*. Baltimore: Johns Hopkins Press.

Carey, Mathew (1760–1839)

Henry W. Spiegel

American publicist. Carey came to America as a poor immigrant from Ireland. He settled in Philadelphia, where in 1785 he founded a publishing, printing and bookselling business that eventually became the largest of its kind in the United States; a successor firm is still in the publishing business. Carey became a leading citizen of Philadelphia, got involved in politics, and participated in many local and regional controversies. When, after the end of the War of 1812, the Pennsylvania manufacturers were threatened by a flood of imports, Carey became a leader of the protectionist movement. A prolific writer, he supported its cause by a flood of publications that reached a wide public and helped to establish Hamilton's 'American System'.

In his *Olive Branch* (1814), Carey attempted to reconcile the Federalists and Democrats. A statement promoting protectionism was inserted in later editions of the work, which embodied Carey's message in over 10,000 copies – according to Carey himself, a record for a book not religious in nature.

Among the many pamphlets that Carey wrote in support of various causes, some thirty contain philanthropic appeals aiming to improve the wages and working conditions of the poor. An example of these is his *Address to the Wealthy of the Land* (1831). The free-trade economists with whom he had battled for so long he now takes to task for allegedly discouraging aid to the poor. People, he argues, may be unemployed or casually employed against their wishes, and some work in employments where their supply is large relative to the demand for their labour. He proposes to arouse public opinion against employers who fail to pay a living wage, and points to education and increased mobility of labour as means to improve the position of the poor. Ideas such as these are now commonplaces, but when Carey wrote about them, his was a lonely voice.

References

Nuesse, C.J. 1945. *The social thought of American catholics*. Washington, DC: The Catholic University of America Press.

Rowe, K.W. 1933. *Mathew Carey: A study in American economic development*. Baltimore: Johns Hopkins Press.

Spiegel, H.W. 1960. *The rise of American economic thought*. Philadelphia: Chilton. ch. 5.

Carlyle, Thomas (1795–1881)

Murray Milgate

Keywords

Carlyle, T.; Cash nexus; Democracy; Engels, F.; McCullough, J. R.; Marx, K. H.; Ruskin, J.; Utilitarianism

JEL Classifications

B31

The eldest of nine children of Margaret Aitkin and James Carlyle, Thomas Carlyle was born at Ecclefechan in Scotland on 4 December 1795. While Carlyle's contributions ranged over many fields (including history, literary and social criticism, biography, translation and political commentary), in economics he is remembered chiefly as the originator of the epithet 'the dismal science' ('The Nigger Question', 1849; in *Miscellaneous Essays*, vol. 7, p. 84). Among 'the professors of the dismal science', one M'Croudy (J.R. McCulloch) is a principal target of Carlyle's criticism. Yet Carlyle's writings on economics are more extensive than this small measure of recognition might suggest, and his key criticisms of the economic and political tendencies of the 'present times' (as he called them) are contained essentially in three works: *Chartism,* (1840), *Past and Present* (1843) and *Latter-Day Pamphlets* (1850). Almost inevitably, Carlyle's characteristically romantic reaction to the decline of authority and the rise of utilitarian individualism led him into head-on collision with the prevailing economic doctrines of the day. Since, for Carlyle, the challenge of democracy to the *ancien régime* had been carried forward under the mistaken banner 'Abolish it, let there henceforth be no relation at all' (1850, p. 21), it was natural for him to hold that laissez-faire, free competition, the law of supply and demand, and the 'cash nexus' were no more than 'superficial speculations . . . to persuade ourselves . . . to dispense with governing' (1850, p. 20). Although Carlyle's account of the 'cash-nexus' was adopted verbatim by Marx and Engels in the opening pages of *The Communist Manifesto*, in the latter sections of that document his overall position is roundly attacked (see there the reference to the 'Young England', of which Carlyle was a prominent member).

There is also a thinly veiled attack on Carlyle's 'dissatisfaction with the Present . . . and affection and regret towards the Past' in John Stuart Mill's *Political Economy* (1848, pp. 753–4). However, at Carlyle's hands the utilitarian calculus of pleasure and pain fared little better. It was charged

with ignoring all those sentiments, aspirations and interests which distinguished the human from other animals and was dubbed by Carlyle 'the Pig Philosophy' (1850, p. 268). Though Carlyle had few if any followers among economists, he exerted a profound impact upon the thinking of John Ruskin, and he may correctly be regarded as a principal exemplar in England of that reactionary or feudal brand of 'socialism' criticized by Marx and Engels in the *Communist Manifesto.* Carlyle died in Chelsea on 5 February 1881 and was buried in Ecclefechan.

Selected Works

1888–9. *Works*, 37 vols. London: Chapman & Hall. (Page references above are from this edition.)
1896–9. *Works.* The centenary edition in 30 vols. London: Chapman & Hall.

Bibliography

Mill, J.S. 1848. *Principles of political economy*, ed. W.J. Ashley from the 7th edn (1871). London: Longmans, 1909.

Carroll, Lewis (Charles Lutwidge Dodgson) (1832–1898)

Bernard Grofman

Born on 27 January 1832, he was Student at Christ Church, Oxford, 1852–98, and Lecturer in Mathematics 1856–81. He died on 14 January 1898.

Lewis Carroll was the author of *Alice's Adventures in Wonderland* (1865), *Through the Looking Glass and What Alice Found There* (1872), and a large number of humorous poems of which 'The Hunting of the Snark' (1876) is the best known. In his real identity, that of Charles L. Dodgson, he was a mathematician of modest repute in the areas of geometry, recreational mathematics, and logic: author of *Euclid and his Modern Rivals* (1879), *Curiosa Mathematica* (1888, 1893), and *Symbolic Logic*, *Vol. I* (1896). Under either identity, however, he may appear to be a rather unlikely candidate for inclusion in an encyclopedia of economics. Yet his work on mechanisms for political representation anticipates important ideas in game theory and that branch of public choice theory having to do with committees and elections. The earliest work appeared in three privately printed pamphlets on *The Theory of the Committee* (1873, 1874, 1876) and dealt with a number of topics in majority rule procedures including a discussion of what is known today as the Borda count. Only recently has it been rediscovered and the significance of its contributions realized – almost entirely because of the historical scholarship of Duncan Black (1958, 1967, 1969, 1970).

The Principles of Parliamentary Representation (1st edn, Nov. 1884, 2nd edn, Jan. 1885), applies techniques which we now associate with two-person zero-sum games to solve the problem of the optimal strategy for a two-party competition in a class of voting games in which each party must decide how many candidates it wishes to nominate in a constituency in which each voter may cast v votes (no more than one to each candidate) and there are m seats to be filled. If $v < m$-1 we have what is called the limited vote. If $v = m$ we have plurality or the bloc vote. To make the problem tractable, Dodgson supposes that each of the parties knows the number of its own supporters and those of the opposing party and that each party is able to direct the voting of each of its supporters exactly as it chooses. While not, of course, referring to it as such, he makes use of the idea of a maximin strategy in which each party chooses under the assumption that the opposing party will be optimally distributing its voting strength among an optimal number of candidates.

In this same work, Dodgson considers the question of what voting rule of the type specified above will be optimal in the sense of minimizing the expected proportion of voters whose votes are 'wasted'. By a 'wasted' vote Dodgson here means that the voter's ballot played no part in effecting the outcome; e.g. if a party with s per cent of the

electorate elects h candidates but would have elected that same number of candidates even if it had received support from only s' percent of the electorate ($s' < s$), then ($s - s'$) per cent of the electorate has had its votes wasted. In Dodgson's view, the existence of wasted votes implies that some voters are not having their preferences fully represented. He finds $v = 1$, a special form of the limited vote, commonly called the single non-transferable vote (used in post-World War II Japan) to be optimal under this standard. Under the assumption of a rectangular distribution of party voting support, he finds that the reduction in the magnitude of the expected wasted vote drops off rapidly with increasing m, for $m > 4$.

In related work, Dodgson uses a game-theoretic style of argument to consider optimal party candidate strategies under a cumulative voting system (a semi-proportional system in which each voter may cumulate up to v votes on a single candidate) and under the Hare system (the single transferable vote, a proportional system in which voters indicate their relative orderings of the candidate). For the latter election system, Dodgson looks at the problem of rational coalition forming and provides some examples to show that the results of the Hare system need not be consistent with the expected outcome of a coalitional bargaining game between political parties. However, Dodgson's results are at best suggestive. Indeed the problem he posed has only just been solved (Sugden 1983).

Dodgson's work on proportional representation was guided by his familiarity with research done by a number of Cambridge mathematicians (most involved to some degree with the Proportional Representation Society), a group whom Black (1970) identifies as the Cambridge School of Mathematical Politics. While Dodgson's treatment of proportional representation takes some essential ingredients from these earlier writers, his systematic treatment of the limited vote is a new creation. 'Where there had been only scattered fragments, he leaves a completed edifice' (Black 1970). In making use of the maximin strategy to obtain an equilibrium solution to a particular two-person zero sum game and in examining optimal coalitional strategies in the context of election politics, Dodgson's long-neglected work deserves recognition as a step on the road toward the development of the modern theory of political economy.

See Also

- ▶ Black, Duncan (1908–1991)
- ▶ Borda, Jean-Charles de (1733–1799)
- ▶ Voting

Selected Works

1865, 1872. Carroll, Lewis. *The annotated alice: Alice's adventures in wonderland and through the looking glass,* ed. Martin Gardner. New York: New American Library, 1960.

1873, 1874, 1876. Dodgson, Charles L. *The theory of the committee*. Privately printed.

1884, 1885. Dodgson, C.L. *The principles of parliamentary representation*. London: Harrison.

1888, 1893. Dodgson, C.L. *Curiosa Mathematica*. London: Macmillan.

1879. Dodgson, C.L. *Euclid and his modern rivals*. London: Macmillan.

1896. Dodgson, C.L. *Symbolic logic, vol. I*. Reprinted, along with *Symbolic logic, vol. II*, ed. William W. Bartley, III. London: Macmillan.

Bibliography

Black, D. 1958. *The theory of committees and elections*. Cambridge: Cambridge University Press.

Black, D. 1967. The central argument in Lewis Carroll's 'The principles of parliamentary representation'. *Papers on Non-Market Decision Making*, Fall.

Black, D. 1969. Lewis Carroll and the theory of games. *American Economic Review Proceedings* 59(2): 206–216.

Black, D. 1970. Lewis Carroll and the Cambridge mathematical school of P.R.: Arthur Cohen and Edith Denman. *Public Choice* 8: 1–28.

Black, D. (forthcoming) *Lewis Carroll*. Lennon, F.B. 1962. *The life of Lewis Carroll*. New York: Macmillan.

Phillips, R. (ed.). 1971. *Aspects of Alice*. New York/London: Vanguard/Gollancz. reprinted, New York: Vintage, 1977.

Sugden, R. 1983. Free association and the theory of proportional representation. *American Political Science Review* 78(1): 31–43.

Cartel

Leonard W. Weiss

A cartel, according to Webster, can be either 'a written agreement between belligerent nations' such as a prisoner exchange arrangement, or 'a voluntary, often international combination of independent private enterprises supplying like commodities or services' (Webster's 1967). The second concept is our concern here. The majority of cartels have dealt with national or smaller markets, but many of the best known have been international in coverage. Economists often distinguish private cartels and public cartels. In the latter, the government theoretically makes the rules, typically under strong influence from the affected industry and enforces them. Private cartels involve private agreements. They may or may not be publicly enforced depending on the nation, the period and the agreement. Some international cartels are private, but the best known have resulted from agreements among national governments.

Cartels may involve price fixing, output controls, bid-rigging, allocation of customers, allocation of sales by product or territory, establishment of trade practices, common sales agencies or combinations of these. Many medieval cities and mercantilist nations were tightly bound by such restraints of trade, but the cartel movement is usually pictured as arising with the large private firm in the late nineteenth century. Cartels were carried farthest in Germany in the half-century ending with World War II, but they were also important in Austria, Switzerland, Italy, France, Scandinavia and Japan in the same period. They reached their peak during the great depression of the 1930s. Cartelization was slower to develop in Britain and other nations with a common law tradition such as the United States. A prohibition of contracts in restraint of trade (largely a refusal of the courts to enforce) goes back at least to the early fifteenth century in English common law. The prohibition was written into the American Sherman Anti-Trust Act when it was passed in 1890. Even in the United States, however, the National Industrial Recovery Act, passed at the bottom of the great depression in 1933, permitted industries to formulate enforceable 'codes of fair competition'. The Act was ruled unconstitutional by the Supreme Court in 1935, but the United States continued public cartels in such fields as coal-mining, oil production, interstate transportation, and agriculture for many years.

In the years since World War II most private and public industrial cartels have weakened. America's prohibition of private cartels was strengthened and many of its public cartels ended. The Western occupation forces in Japan and Germany imposed cartel prohibitions there. The subsequent national governments revised these rules to permit certain cartels, but they are far removed from the prewar, pro-cartel policies of the same countries. Most other industrial non-communist countries have adopted anti-cartel laws since the war, but few have gone as far as the United States. On the other hand, some international commodity agreements established extremely high prices in the 1970s. In a number of cases such as bauxite and copper the agreements failed within a few years. But the Organization of Petroleum Exporting Countries (OPEC) was able to keep world oil prices far above their costs for more than a decade. This was possible because Saudi Arabia, with more than a quarter of world capacity, was willing to reduce output greatly as smaller producers inside and outside OPEC expanded.

A 'perfect cartel' is one that maximizes the sum of the profits of its members. This requires that output be allocated among participants so that cost is minimized. That, in turn, implies that different producers operate their capacities at different rates. In the long run, some participants' plants would be closed. The traditional solution for private cartels is side payments from the expanding to the contracting producers. In fact, although such payments have been made, perfect cartels have generally been beyond the reach of private cartels short of merger. A perfect cartel would be difficult to distinguish from a well-run firm. The classic example was the prewar German chemical

firm, I.G. Farben (Interessen Gemeinschaft Farbenindustrie meaning 'Community of Interests in the Dye Industry'). It did begin as an eight-firm cartel, but by 1925 they had all merged (Michels 1928).

Enforcement is a crucial aspect of cartels. This requires (a) detection of violations and (b) sanctions on violators. Detection is easy in oral auctions. Violations are immediately obvious when they occur. In the more common cases where firms must bid for customers in sealed-bid auctions or through salesmen, detection is much more difficult, unless winning bids are publicly announced. Cumulative changes in market shares seem to be the most credible evidence of whether 'cheating' is going on or not, but the usefulness of such evidence rapidly declines as the numbers of competing firms increase (Stigler 1964). The implication is that purely private cartels with many members are weak. If they have serious social cost it is most likely to work via changes in institutions such as the establishment of a basing point system or political pressure for oral auctions. Public enforcement seems essential for cartels to raise price for any length of time on unconcentrated markets.

Private enforcement also requires privately imposed sanctions. Oral auctions help here also. Only one conspirator per bid need incur any risk in punishing a violator, and even he need not always 'win'. Punishment in these cases is not severe. Since the violator is usually free to withdraw from the bidding, he need not pay a price that involves a loss. He can be deprived of the gain from collusion and, perhaps, access to the objects being bid for. With sealed bids or where the rivals solicit customers through salesmen, punishment is apt to mean general price wars – the temporary suspension of the cartel. As the numbers in a cartel grow, the gain from violating it generally increases faster than the loss from such punishment when detected (Lambson 1984). Here is another reason to expect purely private cartels to be weak unless the market is concentrated.

In private cartels prices are unlikely to be set at joint maximizing levels. The bargaining power of major participants is apt to reflect their potential profitability without the cartel. Usually the low-cost firms have the best prospects without the cartel. If they determine cartel price, it is likely to be lower than that of a monopolist with the same plants. Small firms may also have a special influence on cartel price. A firm that is too small to be worth disciplining will probably sell at a discount from cartel price. Such a small firm as a cartel member is apt to favour high cartel prices from which it then discounts. If the numbers of such small firms become large, the majors may try to discipline the fringe as a whole to limit their discounts. In fact, however, the growth of a large fringe commonly leads to the collapse of the cartel.

Public cartels are also unlikely to be perfect cartels, but they often differ from private cartels. Many American public cartels (such as those in agriculture, oil prorationing, import quotas for oil refiners, and airlines under the Civil Aeronautics Board) allocated output, access to cheap imports of oil or to profitable markets in favour of small and high-cost firms, just the opposite of what would have occurred under successful private cartels.

Effective cartels are likely to result in excess capacity for several reasons. High-profit prospects attract entrants – as in American oil prorationing in the 1940s, 1950s and early 1960s or the famous oil glut that grew up in the 1980s after the huge price increases imposed by OPEC in the 1970s. High prices permit continued excess capacity that would be driven from the field in a competitive market – as in much of American agriculture. Existing firms will often build excess capacity if it increases sales because with prices far above marginal cost, additional sales are worth the additional cost to the firm (Posner 1975) – as occurred on competitive airline routes in 1945–1977 even though entry was prohibited (Douglas and Miller 1974). An equilibrium at high cartel prices is reached when excess capacity has forced cost up to the point where profits are reduced to normal levels and entry and expansion is no longer attractive. Excess capacity can be avoided if members' output does not depend on current capacity. For instance, American flue-cured tobacco production depends on acreage allotments set in the late 1930s. As a result, the government was able over

many years to prevent the development of excess capacity which presented such problems in other crop programmes.

Excess capacity may arise in private cartels also. In addition to entry and expansion attracted by high prices, excess capacity may be intentionally built or maintained so that the threat of retaliation against violators of the cartel can be credible (Brock and Scheinkman 1985).

Many nations have permitted and enforced cartels of certain sorts which were seen to be in the public interest. Most of the industrial non-communist countries of the world including the United States permit export cartels. From a narrow national point of view this makes some sense at least for large countries. Their national incomes are enhanced by exploitation of any powerful positions they occupy abroad. With all major countries following such strategies, however, the overall effect must be some net loss for most of them. Another reason for export cartels arises when a major importing country negotiates a restriction on exports from foreign sources. The Americans negotiated many such quotas with major foreign exporters to the United States in the 1960s and 1970s.

Import quotas almost always involve public cartels in form. Import licences are distributed among importers by the government and are kept valuable by the trade restriction itself. In the 1930s Germany used import restrictions along with complex foreign-exchange policies to exploit its special position with respect to many of its trading partners. The main purposes of import quotas today are protectionism and/or the allocation of scarce foreign exchange. Exploitative import cartels seem to be few. The large countries employ import quotas very little today, and small nations have little monopsony power. Because of international specialization, small countries can often be large in their main export markets, but specialization in consumption and imports is rare.

A number of countries use cartels to aid temporarily depressed industries. The Japanese 'depression cartels' are an example (Hadley 1970). Depressed industries can form cartels for 1 year or less if approved by a specified government agency. The state of the industry need not derive from a general depression, but the case for such cartels seems strongest in such a setting. No long-term adjustment by the industry is called for, and a temporary cartel may be one of the less costly ways of assisting industries seriously hurt by general economic decline. In normal times occasional bankruptcies may serve to weed out badly managed firms, and economic pressure on a declining industry serves to transfer resources to more productive uses, but widespread financial disasters during a depression seem of little social value. The crucial thing is that the depression cartel be truly temporary and that the problems that made the industry 'depressed' do not call for long-term adjustments.

Japanese cartel law also provides for 'rationalization cartels' (Hadley 1970), which are not so limited in duration as the depression cartels. They require the approval of the appropriate public agency, once more. A number of European nations also provide for rationalization cartels. Rationalization refers to long-term adjustments by an industry such as the replacement of suboptimal or obsolete capacity or the elimination of excess capacity. It is conceivable that joint action by the firms in an industry could offer a better solution to excess capacity than a fight to the finish on the open market might yield. At least the transition would be less painful if a joint decision were made about which plants should be closed and the survivors bought out the firms which were to go out of business. In practice, rationalization cartels have done little of this. Rather, they set price and/or output that reduced the pressure on their members to adjust. They accomplished little or no rationalization as a result.

Where rationalization means replacing suboptimal or obsolete capacity, the cartel approach seems even less promising. It would call upon efficient producers to help their high-cost rivals to become more competitive. A theoretically appealing exception is the specialization cartel. The firms in such a cartel agree to assign products to particular members, thus permitting optimal-scale capacity for each subproduct. Governments that permit such cartels often try to reduce their competitive effects by limiting the combined

shares of the market of the cartel. For instance, in the European Coal and Steel Community such cartels may not have more than 15% of industry sales. However, four such groups permitted in Germany, each with a common sales agency, accounted for most of German steel and half of ECSC steel in the 1960s. Most of the specialization involved output quotas which permitted economies of long production runs. Little specialization of plant and equipment was accomplished, so few economies of scale were realized (Stegemann 1979).

In general, most rationalization cartels have turned out in fact to be oriented primarily toward short-term restraint of trade.

See Also

- ▶ Anti-trust Policy
- ▶ Collusion
- ▶ Cooperative Equilibrium
- ▶ Industrial Organization
- ▶ Market Structure
- ▶ Monopoly
- ▶ Oligopoly
- ▶ Rationalization of Industry

References

Brock, W.A., and J. Scheinkman. 1985. Price setting supergames with capacity constraints. *Review of Economic Studies* 52(3): 371–382.

Douglas, G.W., and J.C. Miller III. 1974. *Economic regulation of domestic air transport*. Washington, DC: Brookings.

Fuller, J.G. 1962. *The gentlemen conspirators*. New York: Grove (About the American electrical equipment cartel of the 1940s and 1950s).

Hadley, E.M. 1970. *Anti-trust in Japan*. Princeton: Princeton University Press. ch. 15.

Lambson, V. 1984. Self-enforcing collusion in large dynamic markets. *Journal of Economic Theory* 34(2): 282–291.

MacAvoy, P.W. 1965. *The economic effects of regulation*. Cambridge, MA: MIT Press.

Michels, R.K. 1928. *Cartels, combines, and trusts in post-war Germany*. New York/London: Columbia University Press/P.S. King & Co.

Osborn, D.K. 1976. Cartel problems. *American Economic Review* 66: 835–844.

Posner, R.A. 1975. The social cost of monopoly and regulation. *Journal of Political Economy* 83(3): 807–827.

Pribram, K. 1935. *Cartel problems*. Washington, DC: Brookings.

Stegemann, K. 1979. The European experience with exempting specialization agreements and recent proposals to amend the Combines Investigation Act. In *Canadian competition policy*, ed. J.R.S. Prichard, W.T. Stanbury, and T.A. Wilson, 449–486. Toronto: Butterworths.

Stigler, G. 1964. A theory of oligopoly. *Journal of Political Economy* 72(1): 44–61.

Stocking, G.W., and M.W. Watkins. 1946. *Cartels in action*. New York: The Twentieth Century Fund. *Webster's Unabridged International Dictionary*. 1967. 3rd edn, Spring-field: Meriam–Webster.

Cartels

Margaret C. Levenstein and Valerie Y. Suslow

Abstract

Cartels are associations of firms that restrict output or set prices. They may divide markets geographically, allocate customers, rig bids at auctions, or restrict non-price terms. They have often been formed with the participation or support of state actors. In contrast to the pre-Second World War period, today most cartels are illegal in most jurisdictions. The average duration of cartels is between five and 7 years, but the distribution of duration is skewed: a large number of cartels break down within a year but a sizable proportion last for over a decade.

Keywords

Antitrust enforcement; Barriers to entry; Cartels; Cheating; Collusion; Communication; Concentration; Coordination; Entry; Innovation; Price fixing; Price wars; Productivity growth; Trust

JEL Classifications

L13

Producers form cartels with the goal of limiting competition to increase profits.

Cartels are associations of independent firms that restrict output or set prices. They may divide markets geographically, allocate customers to specific producers, rig bids at auctions, or restrict non-price terms offered to customers. They have often been formed with the active participation or support of state actors. In contrast to the pre-Second World War period, today most cartels are illegal in most jurisdictions.

Upon its creation a cartel immediately faces three key problems: coordination, cheating and entry. In a dynamic economy, the solution to these problems will change over time, so successful cartels must develop an organizational structure that allows them to re-solve these problems continuously.

Stigler's (1964) classic article highlights the incentive to cheat as the most important source of instability undermining cartels. In a repeated setting, a firm weighs the expected gain from cheating today (the benefit from cheating) with the expected reduction in future discounted profits that follows cheating (the cost of cheating). In order for firms to be willing to refrain from cheating, the following must hold:

$$\sum_{t=0}^{T} \delta^t \frac{\prod^m}{n} \succ \Pi^m$$

where Π^m is the one-period cartel profit, n is the number of firms in the industry, and δ^t is the discount rate. Thus, collusion is easier to achieve the larger the difference between cartel and non-cartel profits, the smaller the number of firms, and the more patient these firms are (Tirole 1988).

Friedman (1971) demonstrates that firms may use 'off the equilibrium path' threats of price wars in retaliation for cheating to provide firms with the incentive not to cheat. However, because in his model any cheating would be observed immediately and therefore subject to swift retaliation, firms do not cheat and price wars are not observed. In the Green and Porter class of models (Green and Porter 1984; Abreu et al. 1986), firms cannot observe one another's output (or pricing) actions nor infer them with certainty from public information. Economic fluctuations require that firms revert to equilibrium 'punishment' or 'price war' behaviour at times in order to maintain the incentives necessary to achieve collusion. Thus, the appearance of on-and-off collusion does not represent inherent cartel instability, but rather a mechanism that cartels use to stabilize themselves.

This theoretical perspective also implies a second mechanism for increasing cartel stability: a cartel may invest in information collection in order to better monitor individual firm activities. Improved monitoring both deters cheating and allows cartels to avoid costly price wars that arise from the inability to distinguish cheating from external shocks.

The most successful cartels actively work to create barriers to entry. Sometimes this is done through collective predation, as in Scott Morton (1997) in which incumbent cartel members successfully deterred entry by financially weaker and smaller firms. In other cases, cartels have turned to the state to create regulations, tariffs, or provide anti-dumping protection with the goal of excluding outsiders. Cartels sometimes use vertical exclusion (for example, a joint sales agency) or restrict access to technology (for example, via a patent pool) to limit entry.

Cartels use direct and repeated communication to overcome obstacles to coordination. Cartel negotiations often begin with discussions of prices and market shares, but expand over time to restrict cheating in non-price dimensions, such as terms of sale, advertising, transport costs, and production capacities. Firm asymmetries and changes in firms' costs can make these negotiations challenging. Slade (1989) suggests that price wars arise from changes in firm or industry characteristics. These price wars then facilitate the learning necessary for firms to re-establish collusion. Cartels also learn how to structure incentives so that collusion is more profitable in the long run than cheating. For example, successful cartels often fashion self-imposed penalties or other compensation schemes for firms that exceed cartel quotas. Cartels sometimes develop elaborate

internal hierarchies allowing for communication at various levels of management. A hierarchical cartel structure allows for high-level information exchange and bargaining activities to be separated from regional or local information exchange and monitoring efforts. When trust is particularly difficult to establish and firms doubt the accuracy of communication or data exchanges, cartels often turn to a third party – such as a trade association – to facilitate information sharing.

The average duration of cartels measured over a range of countries and time periods is between five and seven years (Levenstein and Suslow 2006). There is considerable dispersion in cartel duration: the standard deviation of duration is almost as high as the average. Observed cartel duration is very skewed, with a large number of cartels lasting less than a year or two and a long tail of cartels that endure for a decade or more.

Predictable fluctuations in product or industry demand do not generally undermine effective cartels, but rapid industry growth and unexpected shocks do. Macroeconomic fluctuations, which are close to common knowledge, have little impact on cartel stability. Many successful cartels develop an organizational structure that allows them to weather cyclical fluctuations. Cartels that are disrupted by observable cyclical fluctuations may be inherently fragile.

Large customers can undermine cartel stability by increasing the incentive to cheat, as posited by Stigler (1964) and tested by Dick (1996). On the other hand, large customers sometimes benefit from the existence of a cartel if they receive preferential pricing compared with that received by their smaller competitors, and can even contribute to its stability.

Although posited by theory, there is no simple empirical relationship between industry concentration and the likelihood of collusion. This may reflect sampling bias in studies that focus on prosecuted cartels, since cartels with many firms or with the involvement of an industry association may be easier to detect. Or it may be that industries with a small number of firms are able to collude tacitly without resorting to explicit cartels. Finally, it may reflect the endogeneity of concentration: collusion may allow more firms to survive and remain in the market (Sutton 1991; Symeonidis 2002).

Analyses of the impact of cartels on prices and profits generally use one of three approaches: changes in price following cartel formation, comparison between 'good times' and 'price war' periods, and, comparison between the cartel price and a counterfactual or 'but-for' price that would have prevailed in the absence of collusion. Connor and Lande (2005) provide an exhaustive survey of studies of cartel price effects. They conclude that the median overcharge resulting from cartels is approximately 25 per cent.

Cartels can also affect investment and productivity. Cartel participants have often argued that cartels increase investment and productivity growth by allowing firms to smooth production over time. Others have argued that, by removing the pressure of competition, cartels reduce innovation and productivity growth. Theoretical models have suggested that cartels lead to increased investment in capacity either because excess capacity can deter entry and provide enforcement (Dixit 1980) or because, when price competition is suppressed, firms compete in other dimensions (Feuerstein and Gersbach 2003). In some cases, cartels explicitly restrict investment in new capacity. Where there are not such explicit restrictions, empirical studies have found cartels are associated with increases in investment. On the other hand, no consistent relationship between cartels and productivity growth or innovation has been established empirically (Symeonidis 2002).

As firms have become increasingly global, international antitrust law and policy has faced new challenges. Competition authorities have increased enforcement, attempted to harmonize practices and procedures, and increased cooperation across jurisdictions. The United States is the country with the longest history of prosecuting explicit collusion, with state laws antedating the national ban on price fixing enacted with the passage of the Sherman Act of 1890. Many Western European countries adopted laws against price fixing following the Second World War, but also allowed a large number of exemptions. Since the mid-1990s these exemptions have been sharply reduced, and dozens of other countries have

banned price fixing for the first time. Enforcement activities against cartels, and international cartels in particular, rose sharply in the United States in the late 1990s. European countries, including the newest members of the European Union, have also increased their enforcement activities against cartels, as have countries in Asia, Africa and Latin America. Price fixing – long a criminal offence in the United States – has now been criminalized in several other countries, including the United Kingdom and Ireland. This increased enforcement has demonstrated that cartels continue to be active in a wide range of industries in the 21st century.

See Also

- ▶ Anti-trust Enforcement
- ▶ Cooperation
- ▶ Market Structure
- ▶ Organization of the Petroleum Exporting Countries (OPEC)
- ▶ Stigler, George Joseph (1911–1991)

Bibliography

Abreu, D., D.G. Pearce, and E. Stacchetti. 1986. Optimal cartel equilibria with imperfect monitoring. *Journal of Economic Theory* 39: 251–269.

Connor, J.M., and R.H. Lande. 2005. How high do cartels raise prices? Implications for optimal cartel fines. *Tulane Law Review* 80: 513–570.

Dick, A.R. 1996. When are cartels stable contracts? *Journal of Law and Economics* 39: 241–283.

Dixit, A. 1980. The role of investment in entry-deterrence. *Economic Journal* 90: 95–106.

Feuerstein, S., and H. Gersbach. 2003. Is capital a collusion device? *Economic Theory* 21: 133–154.

Friedman, J.W. 1971. A non-cooperative equilibrium for supergames. *Review of Economic Studies* 38: 1–12.

Green, E.J., and R.H. Porter. 1984. Noncooperative collusion under imperfect price information. *Econometrica* 52: 87–100.

Levenstein, M.C., and V.Y. Suslow. 2006. What determines cartel success? *Journal of Economic Literature* 64: 43–95.

Martin, S. 2002. *Advanced industrial economics*. 2nd ed. Oxford: Blackwell.

Posner, R.A. 1970. A statistical study of antitrust enforcement. *Journal of Law and Economics* 13: 365–419.

Scott Morton, F. 1997. Entry and predation: British shipping cartels 1879–1929. *Journal of Economics and Management Strategy* 6: 679–724.

Slade, M.E. 1989. Price wars in price-setting supergames. *Economica* 56: 295–310.

Stigler, G. 1964. A theory of oligopoly. *Journal of Political Economy* 72: 44–61.

Stocking, G.W., and M.W. Watkins. 1949. *Cartels in action*. New York: The Twentieth Century Fund.

Sutton, J. 1991. *Sunk costs and market structure: Price competition, advertising, and the evolution of concentration*. Cambridge, MA: MIT Press.

Symeonidis, G. 2002. *The effects of competition: Cartel policy and the evolution of strategy and structure in british industry*. Cambridge, MA: MIT Press.

Tirole, J. 1988. *The theory of industrial organization*. Cambridge, MA: MIT Press.

Carver, Thomas Nixon (1865–1961)

A. W. Coats

Carver's career exemplifies the blend of scientific economics and popular social science so characteristic of his period. He was born on 25 March 1865 in Kirkville, Iowa. After a disrupted undergraduate education at Iowa Wesleyan and the University of Southern California (AB, 1891), he studied at Johns Hopkins under Richard T. Ely and John Bates Clark, eventually obtaining his PhD at Cornell in 1894. A joint appointment in economics and sociology at Oberlin led to a professorship in political economy at Harvard (1900–32), where for a time he taught the only course in sociology. His principal theoretical work in economics was an extension of Clark's marginalism to a synthesis of abstinence and productivity theories of interest. He also made pioneering contributions to the economics of agriculture and rural sociology, and published several textbooks and numerous magazine articles. Carver's attacks on radicalism and socialism, his forthright advocacy of individualism, thrift and free enterprise, and his insistence on the crucial value of natural resources conservation and social balance, made him a cult figure among Harvard students. Acceptance of Malthusian population theory and recognition of the dangers of corporation power did not quench his optimism, although

he favoured public works and credit expansion as a corrective to the 1930s depression. Carver served as adviser to the Department of Agriculture and Director of its rural organization service in 1913–14. An energetic and successful Secretary-Treasurer of the American Economic Association from 1909 to 1913, he was elected President in 1916. He died in Santa Monica, California, on 8 March 1961.

Selected Works

1893. The place of abstinence in the theory of interest. *Quarterly Journal of Economics* 8:40–61.
1904. *The distribution of wealth*. New York: Macmillan; reprinted, 1932.
1911. *Principles of rural economics*. Boston: Ginn & Co; reprinted, 1932.
1912. *The religion worth having*. Boston: Houghton Mifflin; revised ed., Los Angeles: Ward Ritchie Press, 1940.
1915. *Essays in social justice*. Cambridge, MA: Harvard University Press; repr. 1940.
1919. *Principles of political economy.* Boston: Ginn & Co.
1921. *Principles of national economy.* Boston: Ginn & Co.
1925. *The Present economic revolution in the United States*. Boston: Little, Brown & Co.
1935. *The essential factors of social evolution*. Cambridge, MA: Harvard University Press.
1949. *Recollections of an unplanned life*. Los Angeles: Ward Ritchie Press.

Case-Based Decision Theory

Ani Guerdjikova

Abstract

Case-based decision theory was developed by Gilboa and Schmeidler. This article describes the framework and lays out the axiomatic foundations of the theory. An illustration based on a model of repeated choice is provided and the applications of the theory to economic problems are listed. Finally, the relationship between the case-based decision theory and expected utility theory is discussed.

Keywords

Aspiration levels; Case-based decision theory; Consumer choice theory; Decision making; Expected utility theory

JEL Classification

G11; D81; D83

Case-based decision theory (CBDT), developed by Gilboa and Schmeidler (1995, 1997a, 2001a; Gilboa et al. 2002), models decision situations in which neither states of the world nor their probabilities can naturally be inferred from the description of the problem. Instead, the decision maker (DM) has a memory of cases, recording the outcomes of acts in problems encountered in the past. For a specific problem, the evaluation of an act is given by a weighted sum of the utilities of outcomes observed in the memory. The weights represent the similarity of the problem-act pairs in those cases in which the outcomes occurred to the problem-act pair under consideration.

An Example

Consider a CEO who seeks to hire an administrative assistant. The available acts are the various candidates for the job. The CEO does not know how well each of the candidates would perform if actually hired. Each candidate might turn out to be unreliable, dishonest or incompetent. Some candidates might be very efficient at administrative tasks, but unable to deal with customers. Others might be perfect on the job, but unwilling to travel.

In this example, neither the possible outcomes nor the states of the world are naturally implied by the description of the problem. Any attempt to specify these would require imagining every

possible situation in which different characteristics of the candidate might be relevant and assigning to each such situation and each candidate an outcome. A much more realistic approach to the problem is to ask each candidate for references: that is, for records of past cases in which they have been employed and specific outcomes have been observed. To determine the utility assigned to each candidate, the outcomes observed in these past cases are weighted by their relevance (similarity) to the decision at hand.

The Framework

The set of decision situations is P with $p \in P$ being referred to as a *problem.* The decision maker chooses an *act a* out of a set of *available acts, A*. The set of *outcomes* is R with a representative element *r.* The DM does not know the states of the world, the state-contingent outcomes or their distributions. Instead, she uses her *memory, M*, in which information about *past cases* is stored. A case, c, is a triple consisting of a problem encountered, p, an act chosen in this problem, a, and an experienced outcome, r : $c = (p,a,r)$. The set of all possible cases is $C = P \times A \times R$. The memory M is represented by a function: $M : C \rightarrow \mathbb{Z}_0^+$, which lists the number of occurrences of each case in the memory. The order of occurrence of different cases is not specified. This can reflect the belief that the order of outcome realizations does not matter for the evaluation of acts. Alternatively, the time component might be incorporated in the description of the problem. (This invariance property appears as an axiom in Billot et al. 2005. In this formulation of CBDT, which follows Gilboa et al. 2002, it is implicit in the description of the memory.)

Let M be the set of all possible memories: $\mathbb{M} = \{M : C \rightarrow \mathbb{Z}_0^+\}$. The DM has preferences over acts given the problem she faces and given her memory, $\succsim_{p,M}$.

The *similarity function* quantifies the DM's similarity judgment between the choice of act a' in problem p' observed in the memory and the choice of act a in the problem at hand, p. It captures the idea expressed by Hume (1748) that 'from causes which appear similar we expect similar effects'. The similarity function can be formulated as: $s : (P \times A) \times (P \times A) \rightarrow \mathbb{R}$. For instance, a candidate a applying for a position as an administrative assistant at a magazine (problem p) may present references from her previous occupation with a radio station (problem p'). While the two problems are not identical, they might be considered similar, and hence the case (p', a) would be used to evaluate the candidate for the current position. Distinct candidates can also be considered similar, for example if they have similar qualifications, or have graduated from the same school. Finally, similarity might refer to comparisons between problem-act pairs as opposed to comparisons between individual problems or acts. For instance, hiring a candidate who is proficient in Japanese to report on cultural issues from Japan might be considered similar to hiring a candidate with a degree in economics to manage the finance column.

The Representation

For a given problem p and memory M, act a is preferred to act $a', a \succsim_{p,M} a'$, if and only if $U_{p,M}(a) \geq U_{p,M}(a')$ with:

$$U_{p,M}(a) = \sum_{c \in C} M(c)u(r_c)s((p,a),(p_c,a_c)). \quad (1)$$

Here p_c, a_c and r_c are respectively the problem encountered, the action chosen and the outcome observed in case c, and $u(\cdot) : R \rightarrow \mathbb{R}$ is a utility function over outcomes.

Intuitively, for each case, the DM determines the similarity of the problem-act pair (p_c, a_c) to the current decision (p, a). The utility of r_c is then weighted by the similarity $s((p, a), (p_c, a_c))$ and by the number of occurrences of c in the memory, $M(c)$.

In general, the sums of the similarity values related to two distinct acts a and a' are different: that is, $\Sigma_{c \in C} M(c)s((p,a),(p_c,a_c)) \neq \Sigma_{c \in C} M(c)$

$s((p,a'),(p_c,a_c))$. Hence, the utility function is unique only up to a multiplication by a positive constant, while adding a constant to $u(\cdot)$ will in general change the representation. (This distinguishes the concept of utility used here from the notion of utility in classic consumption theory, where the utility function is unique up to arbitrary monotone transformations, as well as from the von-Neumann-Morgenstern (Bernoulli) utility index, which is unique up to positive affine transformations.)

We can therefore distinguish between positive outcomes, for which $u(r) > 0$, and negative outcomes, with $u(r) < 0$. The former correspond to experiences that the DM would like to repeat, while the latter represent experiences she would rather avoid. $\overline{r} \in R$ is a *neutral outcome* if $u(\overline{r}) = 0$. If all outcomes observed in the memory are neutral, all acts are considered indifferent. The neutral outcomes determine the DM's *aspiration level*, $\overline{u} = u(\overline{r}) = 0$, the minimal level of utility she must obtain in order to be satisfied with her choice. Then (1) can be written as: $U_{p,M}(a) = \sum_{c \in C} M(c)[u(r_c) - \overline{u}]s((p,a),(p_c,a_c))$. The representation is preserved if a positive affine transformation is applied to both $u(\cdot)$ and $\bar{u}$.

In the representation above, the aspiration level is constant and does not depend on the memory. (The axiomatization of Gilboa and Schmeidler (1997a) can accommodate some forms of adaptation of the aspiration level, but this requires the similarity function to be memory-dependent.) Various studies (Jucknat 1937; Festinger 1942; Lewin et al. 1944; McClelland 1958; Atkinson and Litwin 1960; Frey et al. 1993; Easterlin 2003) have demonstrated that aspiration levels vary over time depending on the observed outcomes. Hence, in applications, a process of aspiration adaptation is often introduced and studied (see Section 7).

The representation is preserved under positive affine transformations of the similarity function $s((p, a), \cdot)$, which can be normalized to take on values in the interval [0,1]. The similarity values have therefore a cardinal meaning: for instance, if similarity is derived from some metric on $P \times A$, then equivalent metrics can give rise to distinct similarity functions and to distinct preference relations.

Repeated Choice

The role of the main concepts can be understood by considering the special case of repeated choice, i.e. $P = \{P\}$ (from now on, we omit reference to the problem). Let $A = \{a_1 \ldots a_n\}$. u_i denote the utility realization of a_i if chosen. Let $x \in A^\infty$ denote a sequence of choices. We write $x(t)$ for the act chosen at time t. The memory at time t is given by: $M_t((a_i,u_i),x) = |\{\tau \le t | x(t) = a_i\}|$.

The case-based decision rule implies:

$$x(t+1) \in \arg\max_{a \in A} \sum_{a_i \in A} M_t((a_i,u_i),x)s(a_i,a)u_i.$$

The set of paths consistent with this rule is:

$$X =: \{x \in A^\infty | \text{for all } t \in \mathbb{Z}^+, x(t) \in \arg\max_{a \in A} \sum_{a_i \in A} M_{t-1}((a_i,u_i),x)s(a_i,a)u_i\}$$

Of interest are the limit choice frequencies, $f(a,x) =: \lim_{t\to\infty} \frac{|\{\tau \le t | x(t) = a_i\}|}{t}$.

Assume first that:

$$s(a,a') = \begin{cases} 1 & \text{if } a = a' \\ 0 & \text{else} \end{cases} \qquad (2)$$

The choice of aspiration level influences qualitatively the individual's behaviour. Let $A^+ =: \{a_i \in A | u_i > 0\}$ be the set of acts with positive outcomes. If $A^+ \neq \varnothing$, then, for all $x \in X$, $f(a_i,x) = 1$ for some $a_i \in A^+$. Hence, low aspiration levels lead to satisficing behaviour, or habit formation: an act with a positive (but not necessarily maximal) outcome is chosen forever. In contrast, if $u_i < 0$ for all $i \in \{1 \ldots n\}$, then for all $x \in X$, $x \in X$, $\frac{f(a_i,x)}{f(a_j,x)} = \frac{u_i}{u_i}$; see Gilboa and Schmeidler (2001b). (Analogous results can be derived if the utility realization of a_i is a random variable with mean μ_i and finite variance. In this case, u_i has to be replaced by μ_i: see Gilboa and

Pazgal 2001.) A high aspiration level thus implies switching, or change-seeking behaviour. Acts with higher utility realizations are chosen with higher frequency.

Now consider the impact of similarity. Let $u_i < 0$ for all $i \in \{1 \ldots n\}$, so that change-seeking behaviour is implied. A positive/negative $s(a, a')$ makes the choice of a less or more desirable if a' has been chosen before. If acts are consumption goods, positive/negative similarity can be taken to mean that a and a' are substitutes/complements: see Gilboa and Schmeidler (1997b).

To illustrate how similarity perceptions affect the frequencies of choices, consider $n = 3$ and let $s(a_i, a_i) = 1$ for all $i \in \{1,2,3\}$, $s(a_1, a_2) = s(a_2, a_3) = \overline{s} \in (0, 1), s(a_1, a_3) = 0$. Satisficing occurs if $u_i > 0$ for $i \in \{1,2,3\}$. Let $\overline{s} < \frac{1}{2}$ and $u_i < 0$ for $i \in \{1,2,3\}$. In the limit, all three acts are chosen with positive frequencies: $\frac{f(a_1,x)}{f(a_2,x)} = (1-\overline{s})\frac{u_2}{u_1}, \frac{f(a_3,x)}{f(a_2,x)} = (1-\overline{s})\frac{u_2}{u_3}$ and $\frac{f(a_1,x)}{f(a_3,x)} = \frac{u_3}{u_1}$. Hence, similarity affects the frequencies with which various acts are chosen. The form of the similarity function can also affect the set of acts which are chosen in the limit. For $\overline{s} > \frac{1}{2}, f(a_2, x) = 0, \frac{f(a_1,x)}{f(a_3,x)} = \frac{u_3}{u_1}$ obtains for all x with $x(1) \neq a_2$, independently of the value of u_2, in particular, even if $u_2 > 0$. Although a_2 is never chosen, and *has a positive utility realization*, its evaluation is negative, because of its similarity to other acts with negative utility realizations. (Guerdjikova 2007 generalizes this intuition to a continuum of acts with random utility realizations, and identifies properties of the similarity function which lead to satisficing or switching behaviour.)

Applications

Several applications of the CBDT are related to consumer choice theory: Gilboa and Schmeidler (2001b) relate similarity perceptions to substitutability/complementarity between goods; Gilboa and Schmeidler (1997b) use the CBDT to explain 'brand switching' behaviour, while Gilboa and Pazgal (2001) model the consumer's reaction to price increases.

Aragones (1997) uses CBDT to explain the presence of swing voters in a model of political party competition. Jahnke et al. (2005) analyse a production choice problem. Blonski (1999) models social learning, using different similarity functions to capture differences in social structures. Pazgal (1997) shows that case-based learning can lead to Pareto-optimal outcomes in coordination games. Krause (2009) studies herding behaviour. Gayer (2007) analyses a process of probability perception formation using a similarity function which changes with experience. Guerdjikova (2004, 2006) applies the CBDT to financial markets.

Gilboa et al. (2006, 2009) propose a method for estimating the similarity function from data. Gayer et al. (2007) use this method to test whether housing market data in Tel-Aviv are consistent with case-based optimization, and find that this is indeed the case for the renting segment. Grosskopf et al. (2008) test the CBDT in a laboratory setting and find some evidence supporting the theory.

Axiomatization

Gilboa et al. (2002) provide an axiomatization of (1). For a fixed problem $p \in P$, consider a family of preference relations over acts $\left(\succsim_{p,M}\right)_{M \in \mathbb{M}}$. Hence, the DM has preferences over acts not only for the actually observed memory, but also for any hypothetical memory in $\mathbb{M}$. The *combination* of two memories, M and M' is another memory $M'' \in \mathbb{M}$ defined as $M''(c) = M(c) + M'(c)$ for all $c \in C$

Axiom 1 (Order) For every $M \in \mathbb{M}, \succsim_{p,M}$ is complete and transitive.

Axiom 2 (Combination) If $a \succsim_{p,M} a'$ and $a \succsim_{p,M'} a'$, then $a \succsim_{p,M+M'} a'$.

Axiom 3 (Archimedian) If $a \succsim_{p,M} a'$, then for every $M' \in \mathbb{M}$, there exists $a k \in \mathbb{N}$ such that $a \succsim_{p,kM+M'} a'$.

Axiom 1 is standard and without it a real-valued representation is impossible. Axiom 2 states that if two separate pieces of evidence support the choice of act a more than that of a',

then so should their combination. If a CEO received independent recommendations from two past employers to hire a candidate, he would not need to bring the two employers together for a consultation. Combining the two 'memories' would not change the recommendation. Axiom 2 is less compelling in the context of hypothesis testing: two memories might both be too short in order to reject a given null hypothesis, but the combination of them might contain a sufficient number of observations for the hypothesis to be rejected. Axiom 2 is also violated if similarity perceptions depend on the experience: see Gilboa and Schmeidler (1993, 2003) for examples. Axiom 3 states that every evidence that supports a' more than a can be outweighed by a sufficient number of repetitions of cases that support a more than a'. Axiom 3 is violated if an outcome observed from a given act renders it inferior regardless of any further evidence. For instance, an administrative assistant who has been dishonest once might never be able to find an employment, regardless of how many additional good recommendations she presents.

Axioms 1–3 are consistent with a representation of the following type: $a \succsim_{p,M} a'$ if and only if $\Sigma_{c \in C} M(c) v((p,a),c) \geq \Sigma_{c \in C} M(c) v((p,a),c)$. Here, $u(r_c)s((p, a)(p_c, a_c))$ is substituted with the less informative $v((p, a), c)$. (The axiomatization of this rule is provided in Gilboa and Schmeidler 2001a; 2003).

Let $L : P \times A \to \mathbb{N}$. For $r \in R$, let $L_r \in \mathbb{M}$ denote the memory in which $M(p,a,r) = L_r(p,a)$ and $M(p,a,r') = 0$, for all $r' \neq r$. Hence, each L_r represents a memory in which the only outcome observed is r. The next definition identifies the neutral outcomes in R, i.e. those outcomes for which $u(r) = 0$.

Definition 1 *An outcome $\overline{r}$ is neutral if for every $L : P \times A \to \mathbb{N}$ and every two acts $a, a' \in A, a \succsim_{L_r} a'$.*

Axiom 4 (Diversity) For any four distinct acts, a_1, a_2, a_3 and $a_4 \in A$, and for every non-neutral r $\in R$, there exists an $L : P \times A \to \mathbb{N}$ such that $a_1 \succsim_{p,L_r} a_2 \succsim_{p,L_r} a_3 \succsim_{p,L_r} a_4$. If $|A < 4$, then the same condition holds for any list of length $|A|$.

Axiom 5 (Case-Independence of Desirability) For any r and r' that are non-neutral, either

(i) for every $L : P \times A \to \mathbb{N}$ and all $a, a' \in A, a \succsim_{p,L_r} a'$ holds iff $a \succsim_{p,L_r} a'$, or
(ii) for every $L : P \times A \to \mathbb{N}$ and all $a, a' \in A, a \succsim_{p,L_r} a'$ holds iff $a' \succsim_{p,L_r} a$.

Axiom 4 rules out the case that an act a is dominated by act a' for all possible memories. It precludes, for example, lexicographic preferences of the following type: a CEO working with Japanese clients might feel that it is always better to hire an assistant who speaks fluent Japanese than an assistant who does not, regardless of their letters of recommendation. Finally, Axiom 5 states that the relevance of a case depends only on the problem and the act, but not on the observed outcome. Intuitively, if $a \succsim_{p,L_r} b$, then either outcome r is desirable and the cases in L_r are more similar to (p, a) than to (p, b), or r is undesirable and (p, a) is less similar to the cases in L_r than (p, b). Of course, if cases in the memory are assigned different similarity weights depending on the outcomes observed, this property will not hold.

Axioms 1–5 are sufficient for the existence of the representation and imply its uniqueness in the following sense: if the utility function u and the similarity function s represent $\left(\succsim_{p,M}\right)_{M \in \mathbb{M}}$, then so do αu and $\beta s((p,a),(p',a')) + w_{p',a'}$, where α and $\beta \in \mathbb{R}$ satisfy $\alpha\beta > 0$ and $w_{P',a'} \in \mathbb{R}$ for all $(p', a') \in P \times A$. While Axioms 1–3 and 5 are necessary for the representation, Axiom 4 is not: it imposes an additional linear independence condition on the similarity values for any four distinct acts. However, Gilboa and Schmeidler (2001a) show that Axioms 1–3 and 5 alone are not sufficient, and hence without Axiom 4, a representation of preferences by a real function might not exist.

Case-Based Decision Theory and Expected Utility Theory

We first identify situations in which case-based learning leads to choices that maximize expected

utility with respect to the true distribution of outcomes. In the framework of Section 4, let the utility realization of a_i be a random variable with mean μ_i and finite variance. Denote the aspiration level at time t by $\bar{u}_t$. A path $x = (x_1 = (\overline{u}_1, a_1, u_1) \ldots x_t = (\overline{u}_t, a_t, u_t) \ldots) \in (\mathbb{R} \times A \times \mathbb{R})^{\mathbb{N}}$ describes the aspiration level, the act chosen and the outcomes observed in each period. Let $x^t =: (x_1 \ldots x_t)$.

The case-based rule becomes: $a_{t+1}(x_t) \in \arg\max_{a \in A} \Sigma_{\tau=1}^{t} s(a_\tau, a)[u_\tau - \overline{u}_t]$.

Gilboa and Schmeidler (1996) show that if $s(a, a')$ is defined as in (2) and if for some infinite sparse set $T \subset \mathbb{N}$ and a constant $h \in \mathbb{R}^+$:

$$\overline{u}_t = \begin{cases} \alpha \overline{u}_{t-1} + (1-\alpha) \max\limits_{a \in \{\overline{a} \in A \, s.t. \ |\{\tau | a_r(x) = \overline{a}\}| > 0} \dfrac{\Sigma_{\{\tau | a\tau(x) = a)} u_\tau}{|\{\tau | a_\tau(x) = a\}|} & \text{for } t \notin T \\ \overline{u}_{t-1} + h & \text{for } t \in T \end{cases}$$

then $f\left(x, \text{argmax}_{i \in \{1 \ldots n\}} \mu_i\right) = 1$ obtains almost certainly. The aspiration adaptation process has the property that the DM is realistic, updating his aspirations towards the highest obtained average utility, but also optimistic, increasing her aspirations by $h > 0$ in certain periods. This combination guarantees that her limit behaviour coincides with that of an expected utility maximizer who is informed of the probability distributions over outcomes. (Guerdjikova 2008 extends this result to a more general class of similarity functions. Pazgal 1997 applies the same adaptation rule to strategic interaction and shows that it selects for the Pareto-optimal equilibrium in coordination games.)

In the context of learning from data, Billot et al. (2005) provide a connection between the notion of similarity and the formation of probabilistic beliefs. (Billot et al. 2005 work with a finite set of outcomes containing at least three elements, $|R| \geq 3$. Gilboa et al. 2006 provide an axiomatization for $|R| = 2$, while Gilboa et al. 2009 extend the analysis to the case of a continuously distributed random variable.) Billot et al. (2005) assume that the order in which data arrives is irrelevant. Hence, each data set can be represented by a function $M \in \mathbb{M}$. For a given act a, they consider a mapping $h_a : \mathbb{M} \to \Delta^{|R|-1}$, which associates with each potential memory $M \in \mathbb{M}$ a probability distribution over the possible outcomes of a. The concatenation axiom requires that for all M and $M' \in \mathbb{M}$, there exists an α A (0,1) such that $h_a(M + M') = \alpha h_a\ (M) + (1 - \alpha)\ h_a(M')$. This axiom, together with the requirement that at least three of the vectors $h_a(M)$ are linearly independent, ensures that $h_a(M)$ can be written as $ha(M)(r) = \dfrac{\Sigma_{c \in C^{s(a,c)}} \hat{p}_a^c(r) M(c)}{\Sigma_{c \in C^{s(a,c)M(c)}}}$, where $s(a, c)$ is the perceived similarity between case c and action a, and $\hat{p}_a^c(r)$ is the probability that would have been assigned to outcome r if the memory consisted of only one case c. Hence, probabilities can be represented as similarity-weighted frequencies.

Gilboa and Schmeidler (2001a) emphasize that case-based decision making and expected utility maximization are not rival theories. The linear additive structure of both models implies that we cannot hope to distinguish between the two theories based on their empirical predictions. Matsui (2000) shows the formal equivalence between the two by using a more general formulation of the CBDT, in which similarity also depends on the observed outcomes. His construction shows that embedding CBDT into expected utility theory requires the use of a very large state space, and vice versa, embedding expected utility into CBDT requires a very large set of problems. Hence, the two theories should be viewed as complementary: the language and the concepts of one of the theories will in general appear more suitable for the description of a specific problem than those of the other.

See Also

- ▶ Expected Utility Hypothesis
- ▶ Satisficing
- ▶ Uncertainty

Bibliography

Aragones, E. 1997. Negativity effect and the emergence of ideologies. *Journal of Theoretical Politics* 9: 189–210.

Atkinson, J.W., and G.H. Litwin. 1960. Achievement motive and text anxiety conceived as motive to approach success and motive to avoid failure. *Journal of Abnormal and Social Psychology* 33: 643–658.

Billot, A., I. Gilboa, D. Samet, and D. Schmeidler. 2005. Probabilities as similarity-weighted frequencies. *Econometrica* 73: 1125–1136.

Billot, A., I. Gilboa, and D. Schmeidler. 2008. Axiomatization of an exponential similarity function. *Mathematical Social Sciences* 55: 107–115.

Blonski, M. 1999. Social learning with case-based decision. *Journal of Economic Behavior and Organization* 38: 59–77.

Easterlin, R.A. 2003. *Do aspirations adjust to the level of achievement? A look at the financial and family domains*. Mimeo: University of Southern Carolina.

Festinger, L. 1942. Wish, expectation and group standards as factors influencing level of aspiration. *Journal of Abnormal and Social Psychology* 37: 184–200.

Frey, D., D. Daunenheimer, O. Parge, and J. Haisch. 1993. Die Theorie sozialer Vergleichsprozesse. In *Theorien der Sozialpsychologie I*, ed. D. Frey and M. Irle. Bern: Verlag Hans Huber.

Gayer, G. 2007. *Perception of probabilities in situations of risk a case based approach*. Mimeo: Tel Aviv University.

Gayer, G., I. Gilboa, and O. Lieberman. 2007. Rule-based and case-based reasoning in housing prices. *Berkeley Electronic Press Advances in Theoretical Economics* 7(1), article 10.

Gilboa, I., and A. Pazgal. 2001. Cumulative discrete choice. *Marketing Letters* 12: 118–130.

Gilboa, I., and D. Schmeidler. 1989. Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics* 18: 141–153.

Gilboa, I., and D. Schmeidler. 1993. *Case-based knowledge representation*. Mimeo.

Gilboa, I., and D. Schmeidler. 1995. Case-based decision theory. *Quarterly Journal of Economics* 110: 605–639.

Gilboa, I., and D. Schmeidler. 1996. Case-based optimization. *Games and Economic Behavior* 15: 1–26.

Gilboa, I., and D. Schmeidler. 1997a. Act similarity in case-based decision theory. *Economic Theory* 9: 47–61.

Gilboa, I., and D. Schmeidler. 1997b. Cumulative utility consumer theory. *International Economic Review* 38: 737–761.

Gilboa, I., and D. Schmeidler. 2001a. *A theory of case-based decisions*. Cambridge: Cambridge University Press.

Gilboa, I., and D. Schmeidler. 2001b. Reaction to price changes and aspiration level adjustments. *Review of Economic Design* 6: 215–223.

Gilboa, I., and D. Schmeidler. 2003. Inductive inference: An axiomatic approach. *Econometrica* 71: 1–26.

Gilboa, I., D. Schmeidler, and P. Wakker. 2002. Utility in case-based decision theory. *Journal of Economic Theory* 105: 483–502.

Gilboa, I., O. Lieberman, and D. Schmeidler. 2006. Empirical similarity. *Review of Economics and Statistics* 88: 433–444.

Gilboa, I., O. Lieberman, and D. Schmeidler. 2009. A similarity-based approach to prediction. *Journal of Econometrics*, forthcoming.

Grosskopf, B., R. Sarin, and E. Watson. 2008. *An experiment on case-based decision making*. Mimeo: Texas A&M University.

Guerdjikova, A. 2004. *Evolution of wealth and asset prices in an economy with case-based decision makers*. Discussion Paper 04–49, SFB 504, University of Mannheim.

Guerdjikova, A. 2006. *Asset pricing in an overlapping generations model with case-based decision makers*. Mimeo: Cornell University.

Guerdjikova, A. 2007. Preferences for diversification with similarity considerations. In *Uncertainty and risk: Mental, formal, experimental representations*, ed. M. Abdellaoui, R.D. Luce, M.J. Machina, and B. Munier. Berlin/Heidelberg: Springer.

Guerdjikova, A. 2008. Case-based learning with different similarity functions. *Games and Economic Behavior* 63: 107–132.

Hume, D. 1748. *An enquiry concerning human understanding*. Oxford: Clarendon.

Jahnke, H., A. Chwolka, and D. Simons. 2005. Coordinating service-sensitive demand and capacity by adaptive decision making: An application of case-based decision theory. *Decision Sciences* 36: 1–32.

Jucknat, M. 1937. Leistung, Anspruchsniveau und Selbstbewusstsein. *Psychologische Forschung* 22: 89–179.

Krause, A. 2009. Learning and herding using case-based decision theory. *IEEE Transactions on Systems, Man, and Cybernetics*, Part A, forthcoming.

Lewin, K., T. Dembo, L. Festinger, and P.S. Sears. 1944. Level of aspiration. In *Personality and the behavior disorders*, ed. J. Hunt New. York: Ronald Press.

Matsui, A. 2000. Expected utility and case-based reasoning. *Mathematical Social Sciences* 39: 1–12.

McClelland, D.C. 1958. Risk taking n children with high and low need for achievement. In *Motives in fantasy, action and society*, ed. J.W. Atkinson, 306–321. Princeton: Van Nostrand.

Pazgal, A. 1997. Satisficing leads to cooperation in mutual interests games. *Journal of Game Theory* 26: 439–453.

Cassel, Gustav (1866–1944)

Bo Gustafsson

Keywords

Acceleration principle; Business cycle theory; Cassel, G.; Davidson, D.; Deflation; Hadley, A. T.; Harrod growth formula; Inflation; Kitchin, J.; Labour theory of value; Labour–capital ratio; Marginal cost pricing;

Marginal utility of money; Marginal utility theory of value; Myrdal, G.; Ohlin, B. G.; Purchasing power parity; Quantity theory of money; Say's Law; Spiethoff, A. A. K.; Stockholm School; Trade unions; Tugan-Baranowsky, M. I.; Value; Walras, L.; Wicksell, J. G. K.; Woytinski, W. S.

JEL Classifications
B31

Along with Knut Wicksell and David Davidson, Gustaf Cassel was the founder of modern economics in Sweden. He started as a mathematician and began his career as an economist by treating problems of railway rates and progressive taxation from a mathematical point of view. In order to deepen his understanding of economics he went to Germany, where he attended the seminars of Schönberg, Cohn and other traditional representatives of the economic profession. After visits to England, where he made the acquaintance of Marshall and of Sidney and Beatrice Webb, and a short period of lecturing at the university of Copenhagen, in 1902 Cassel took up a position as associate professor in economics at the university of Stockholm. In 1904 he was appointed a professor in economics and public finance. As holder of the chair he acquired a series of gifted pupils, Gunnar Myrdal and Bertil Ohlin among others, who, although they developed the theoretical heritage of Wicksell rather than that of Cassel, became the founders of the Stockholm School of economics. Before the First World War Cassel frequently served as a government expert on problems of railway rates, taxation, state budgets and banking and his involvement in problems of economic policy increased with the post-war economic problems. During the 1920s he became an adviser to the League of Nations on monetary problems and was commonly regarded as a leading international authority in this field, lecturing and publishing widely. All his life he worked also as a columnist for the Swedish daily paper *Svenska Dagbladet*. Although Cassel was originally liberal, he progressively turned more and more conservative denouncing the labour movement, the welfare state and Keynesianism in the name of 'Modern Scientific Principles'.

It is no easy task to evaluate the contributions of Gustav Cassel to economics. He never cared much about paying homage to his predecessors, from whom he sometimes took over fruitful ideas, while at the same time being unjustifiably critical towards other theorists. His expositions are not seldom marred by contradictions and a vagueness in expression, only scantily veiled by his mastery in round and polished sentences. At the same time Cassel took a keen interest in very many fields of economic theory and practice, he had a firm grip on empirical economics and his gifts in tracking down the relevant and essential aspects of economic problems were unusual. These qualities, in combination with a forceful and pedagogical exposition and, on the top of this, an imperturbable conviction of being the chosen spokesman for progress and the principles of science, made him influential not only among men of practical matters but also among fellow economists.

Cassel's main work is his *Theoretische Sozialökonomie* (1918) but his most important theoretical ideas were in fact conceived already around the turn of the century. In his essay 'Grundsätze für die Bildung der Personentarife auf den Eisenbahnen' (1900b), he criticized the idea of calculating railway rates on the basis of average costs and instead advocated marginal cost pricing. For a railway enterprise as a monopolistic business unit, rates which equalized marginal costs and marginal revenues were the optimal ones, though this might imply that some rates were lower than average costs. Even if the principle had been advocated already in 1885 by the American railway economist A.T. Hadley, it was succinctly formulated by Cassel.

Venturing into general economic theory, Cassel in these years also criticized Ricardo's labour theory of value in the essay 'Die Produktionskostentheorie Ricardos und die ersten Aufgaben der theoretischen Volkswirtschaftslehre' (1901), presented an outline of his own theory of price, 'Grundriss einer elementaren Preislehre' (1899) and developed a theory of interest in *The Nature and Necessity of Interest* (1903). The Ricardian labour theory of value was, according to Cassel, untenable because it

assumed that the labour–capital ratio was equal in different enterprises and industries, that labour was homogeneous and that the marginal land did not pay any rent. He did not care to take issue with the Marxian development of the labour theory of value. The labour theory of value belonged to the so-called one-sided value theories. But so did the marginal utility theory of value, which was deficient primarily because it lacked a clearly conceptualized unit of measurement for utility but also because goods, according to Cassel, are not generally divisible and the valuations of goods are not continuous functions of the supply. Therefore, Cassel suggested that one should do away with all conceptions of value and rest content with money prices and not bother with what might lie behind money prices. Thus Cassel did not consider the fact that money itself may vary in value, nor that the marginal utility of money certainly varies between individuals. Following Marshall, Cassel explained prices by reference to supply and demand and, following Walras, he devised a general equilibrium model for market prices in the form of a system of simultaneous equations. In fact, Cassel's price theory is a simplified version of the theory of Walras, who was characterized as 'in a sense one of my precursors'. However, by popularizing Walras, Cassel contributed much towards the understanding of the mutual interdependencies in a market economy. It was quite logical that the theory of interest that Cassel devised also should be based upon supply and demand, viz. supply of waiting and demand for the use of capital, as a special case of the general theory of price, and he boldly asserted that waiting and use denoted the same thing. Although his theory of interest, showing a close resemblance to that of Senior, was not original, it still merits our attention because of its vivid illustrations and some striking applications. This is particularly the case for Cassel's argument against the idea of a continually falling rate of interest. Given that most saving is made in order to safeguard a permanent future level of income, the shortness of life puts a ceiling under the rate of interest. This was the necessary and sufficient condition for the necessity of interest.

The year after the publication of *The Nature and Necessity of Interest*, Cassel also published his theory of the business cycle and his theory of the secular development of the general level of prices in two articles in the Swedish journal *Ekonomisk tidskrift*, 'Om kriser och däliga tider' (1904a) and 'Om förändringar i den allmänna prisnivän' (1904b). Both these theories were later incorporated and somewhat elaborated in his *Theoretische Sozialökonomie* (1918). In his theory of the business cycle Cassel was evidently influenced by Spiethoff and Tugan-Baranowsky, who recently had made public their theories explaining the business cycle with reference to the variations in investment of fixed capital and of loanable funds. What is really new in Cassel's treatment is his precise formulation of the accelerator principle, which he expounds with reference to the relationship between the demand for freights and the output of ships. The treatment of growth theory had to await the publication of his *Theoretische Sozialökonomie* and also on this point Cassel was wholly original, in fact foreshadowing the Harrod growth formula by his own formula for 'the uniformly progressing economy', the only difference being that Cassel worked with an average instead of a marginal capital coefficient.

Cassel's theory of the secular development of the general level of prices also demands our attention as a piece of brilliant imagination and was as late as 1930, after Kitchin's refinements, accepted as the theoretical basis for the first interim report of the gold delegation of the League of Nations. Cassel's theory was a straightforward quantity theory of money. By calculating the relative variations of gold output in relationship to a calculated normal need of gold for preserving a constant general level of prices, Cassel showed that there was a very good correlation between the relative variations of gold output and the corresponding variations in the general level of prices. Cassel's theory met with all the objections the quantity theory of money usually meets and in addition a series of more specific criticism: that it presupposes a constant ratio between velocity (V) and transactions (T), which is difficult to believe; that it overlooks the important role of silver in the 19th century as well as the varying proportions of the more relevant variable monetary gold; and that a case as good as Cassel's could be made, and in fact was made by Warren and Pearson, by making

the gold price rather than gold output the effective cause of price changes. But since Kitchin's (and Woytinski's) calculations, taking only monetary gold in regard, showed a still better fit between the variations of gold output and prices, Cassel's theory is still a serious candidate.

After this first period of theoretical activity around the turn of the century, Cassel mainly devoted his energy to synthesizing and propagating his ideas on the national and the international scene. The only really new element in his theoretical set-up was the famous purchasing power parity theory of the exchange rates, according to which the international rates of exchanges are determined by the purchasing power of the national currencies. It is easy to show that this is a rather poor general theory for the explanation of the exchange rates. But it contained a pragmatic truth during and after the First World War, when trade balances and, hence, the supply and demand of currencies, to a great extent, were determined by the course of rapid inflation in different countries. It is precisely this instinct for pragmatic truths that explains Cassel's success and influence in the international community of bankers and politicians during the 1920s. In his memoranda to the international conferences of the League of Nations Cassel first and foremost advocated stability of monetary affairs by means of control of the quantity of money, increased interest rates and cut-downs of state expenditures. But he was also critical towards the subsequent ruthless policy of deflation creating widespread unemployment and new disequilibria in world trade as well as intolerable debt burdens. Together with Keynes he criticized the unwillingness of the claimants to the German war debt to receive German goods as payment. When confronted by the permanent unemployment of the 1920s, Cassel concentrated his attacks on trade unions and the level of wages and untiringly explained the gospel contained in Say's Law. During the course of the 1930s it became all too clear that Gustav Cassel had been left behind by the march of events and of economic theory. It was his tragedy that he himself, who once waved his magic wand over international economic affairs, could not bear the truth. After some years of protracted rearguard skirmishes he devoted himself to more philosophical problems and wrote up a voluminous autobiography characteristically entitled 'In the Service of Reason' (*I förnuftets tjänst*, 1940–41). His last words on his death-bed were 'A world currency!'

See Also

- ▶ Purchasing Power Parity
- ▶ Wicksell, Johan Gustav Knut (1851–1926)

Selected Works

1899. Grundriss einer elementaren Preislehre. *Zeitschrift für die gesamte Staatwissenschaften* 55: 395–458.

1900a. *Das Recht auf den vollen Arbeitsertrag.* Göttingen: Vandenhoeck & Ruprecht.

1900b. Grundsätze für die Bildung der Personentarife auf den Eisenbahnen. *Archiv für Eisenbahnwesen*. Trans. as 'The principles of railway rates for passengers', *International Economic Papers* 6 (1956): 126–147.

1901. Die Produktionskostentheorie Ricardos und die ersten Aufgaben der theoretischen Volkswirtschaftslehre. *Zeitschrift für die gesamte Staatswissenschaft* 57: 68–100.

1902a. Der Ausgangspunkt der theoretischen Ökonomie. *Zeitschrift für die gesamte Staatswissenschaft* 58: 668–698.

1902b. *Sozialpolitik.* Stockholm: H. Geber.

1903. *The nature and necessity of interest*. London: Macmillan.

1904a. Om kriser och dåliga tider. *Ekonomisk Tidskrift* 6: 21–35 and 51–81.

1904b. Om förändringar i den allmänna prisnivån. *Ekonomisk Tidskrift* 6: 311–331.

1917. *Dyrtid och sedelöverflöd.* Stockholm: P.A. Norstedt & Söner.

1918. *Theoretische Sozialökonomie.* Leipzig: C.F. Winter. (In manuscript in 1914. There are five German editions.) Trans. into English as *Theory of social economy.* London: T.F. Unwin, 1923; new revised edn., London: E. Benn, 1932. French, Japanese and Swedish editions also available.

1921. *The world's monetary problems. Two memoranda presented to the International*

Financial Conference of the League of Nations in Brussels in 1920 and to the Financial Committee of the League of Nations in September 1921. London: Constable & Co.
1922. *Penningväsendet efter 1914.* Stockholm: P.A. Norstedt. Trans. into English as *Money and foreign exchange after 1914*. London: Constable & Co., 1922. German, French, Spanish and American editions also available.
1925. *Fundamental thoughts in economics.* London: T.F. Unwin.
1927. *Memorandum till den Internationella ekonomiska Konferensen i Genève 1927.* Stockholm: P.A. Norstedt.
1928. *Socialism eller framätskridande.* Stockholm: P.A. Norstedt & Söners.
1935. *On quantitative thinking in economics.* Oxford: Clarendon Press.
1936. *The downfall of the gold standard.* Oxford: Clarendon Press.
1940–41. *I förnuftets tjänst, en ekonomisk självbiograft*, 2 vols. Stockholm: Bokfölaget Natur och Kultur.
1942. *Vår bildnings fåfänglighet.* Stockholm: Bonniers.
1944. *Den odelbara människan.* Stockholm: Natur och Kultur.

Bibliography

A good biography of Cassel has yet to appear. The biography by his secretary, Ingrid Giobel-Lilja (1948), may be consulted for biographical details, especially as regards his family life. Böhm-Bawerk (1914) made a critical examination of Cassel's theory of interest, and Wicksell (1919) published a very critical review of *Theoretische Sozialökonomie.* A more descriptive account of Cassel's main work is to be found in Mitchell (1969); Gunnar Myrdal (1945) wrote an obituary; and Gustafsson (1964) offers another overall evaluation with emphasis on Cassel's theory of long-run prices.
Giobel-Lilja, I. 1948. *Gustav Cassel, En livsskildring*. Stockholm: Natur och Kultur.
Gustafsson, B. 1964. Gustav Cassel (1866–1944). *Vär Tid* No. 3.
Mitchell, W.C. 1969. Chapter 16 in *Types of economic theory,* vol. 2. New York: A.M. Kelley.
Myrdal, G. 1945. Gustav Cassel in memoriam. *Ekonomisk Revy.* Republished (in Swedish) in J. Schumpeter, *Stora nationalekonomer.* Stockholm: Natur och Kultur, 1953.
von Böhm-Bawerk, E. 1914. Exkurs XIII, Kritische Glossen zur Zinstheorie Cassels, Exkurse zur 'Positiven Theorie des Kapitals'.
Wicksell, K. 1919. Professor Cassels nationalekonomiska system. *Ekonomisk Tidskrift* 21: 195–226. Republished in *Schmollers Jahrbuch* 52(5), 1928, and in K. Wicksell, Appendix I in *Lectures on political economy*, vol. I. London: G. Routledge & Sons, 1934.

Caste System

Susan Wolcott

Abstract

India's caste system performed two fundamental functions: insurance through transfers between caste members and, in villages, insurance through protected job assignments across castes. In most of India the landlord had a social responsibility to maintain his lower caste workers in lean periods. This division of labour has been viewed as coercive and exploitative. Yet many groups changed their caste occupation, both upward and downward in ritual ranking. During industrialization, traditional occupational categories did not restrict occupational choices in new industries, but caste continued to play a role in recruitment and support during work stoppages.

Keywords

Akerlof, G; Caste system; Division of labour; India; Industrialization; Insurance

JEL Classification

N3

The caste system in India is a division of society into ranked, hereditary, endogamous occupational groups. It is loosely based on the four *varnas* of Brahmanas (priests), Kshatriyas (warriors and aristocracy), Vaishyas (merchants) and Shudras (the servants of the others). Castes either belonged to one of these four, or were below them in the

hierarchy; these latter are the so-called untouchables. In practice, the *varnas* are less important than were the relationships among and between the numerous sub-castes, or *jatis*. The sub-castes were specific to each region and were the true functional unit of the caste system. They were, for example, the endogamous unit. And obligations of *jati* members to each other were much stronger than were obligations of caste members more generally. Below the terms '*jati*' and 'caste' are used interchangeably.

Caste was not a monolithic institution. Reviewing the historical literature on caste, Rudner (1994, p. 25) notes that it is impossible for any one description to capture the 'on-the-ground diversity of India's caste systems'. He suggests as a definition: 'complex, multilayered, multifunctional corporate kin groups with enduring identities, a variety of rights over property, and crucial economic roles, often within large regions'.

Because of this diversity, caste's role in the Indian economy varied across regions and across groups. But two functions were fundamental: insurance through transfers between caste members and, in village India, insurance through protected job assignments across castes. On the first of these, Srinivas (1962, p. 70) writes, 'joint family and caste provide for an individual in our society some of the benefits which a welfare state provides for him in the industrially advanced countries of the West'. Economists have completely ignored this aspect of caste. But in the modern period it seems to be economically significant: financial transfers among rural villagers are common in developing countries. However, this practice is much more common in India than in any other country yet studied (Cox and Jimenez 1990, Table 1). As caste ties are weakening over time and as income rises, it is likely that such transfers were even more prevalent historically.

And across castes, because each *jati* was, at least in theory, occupationally segregated in the villages of colonial India, it played a protected role in the economic order and had a claim on the wealth produced by the village. This relationship is called the *jajmani* system in much of India, and the *baluta* system in Maharashtra (Kolenda 1978).

A particular division of responsibilities is that between landlords and agricultural labourers. Especially in north, south and east India, the landlord had a social responsibility to maintain his workers in lean periods. Platteau (1995) reviews the literature on this topic and presents a mathematical formalization of this relationship. Greenough (1982) gives an account of the strains on this system and its ultimate collapse in an extreme crisis.

This division of labour has also been viewed as coercive and exploitative. Akerlof (1976) models a situation in which groups can be confined to inferior occupations by social opprobrium. Maddison (1971, p. 28) argues that these occupational divisions were not only coercive but also foolish: 'One might think that some of the lowest productivity occupations were invented simply to provide everyone with a job in a surplus labor situation, but there was no shortage of land and the productivity of the economy would have been higher if there had been greater job mobility.'

But these authors exaggerate the rigidity of the caste system in regard to occupational segregation. Mukerjee (1937) provides a long list of groups which had changed their caste occupation, both upward and downward in ritual ranking, as well as lists of splitting and merging sub-castes. He argues that, although there was rigid social control within the caste, the system revealed 'plasticity' in regard to economic incentives. As an example of this, Commander (1983) notes that historical sources imply that the Chamars of the United Provinces – hereditarily leather workers – were for much of the 19th century largely agricultural labourers. He argues cogently that, although ritual and custom were important in determining economic rewards and relative position in the *jajmani* system, so were land availability and labour scarcity.

Did caste have a role in modern industrialization? The best survey on this subject remains that of Morris (1960). One point is obvious. Traditional occupational categories did not restrict occupational choices in new industries. Whether or not caste affected the economic lives of the workforce in other ways is less clear. Morris (1960, p. 128) writes that he 'is inclined to the

view that *jati* relationships ultimately are irrelevant in the factory'. Most analysts argue, however, that, because of the economically supportive links between *jati* members, caste did have a role in recruitment and support during work stoppages (Chandavarker 1994; Klass 1978). The differentiated and fluid nature of caste makes a general statement impossible.

See Also

- Akerlof, George Arthur (born 1940)
- Labour Market Institutions
- Peasant Economy
- Risk-Coping Strategies

Bibliography

Akerlof, G. 1976. The economics of caste and of the rat race and other woeful tales. *Quarterly Journal of Economics* 90: 599–617.

Chandavarker, R. 1994. *The origins of industrial capitalism in India: Business strategies and the working classes in Bombay, 1900–1940*. Cambridge: Cambridge University Press.

Commander, S. 1983. The jajmani system in north India: An examination of its logic and status across two centuries. *Modern Asian Studies* 17: 283–311.

Cox, D., and E. Jimenez. 1990. Achieving social objectives through private transfers: A review. *World Bank Research Observer* 5: 205–218.

Greenough, P. 1982. *Prosperity and misery in modern Bengal: The Famine of 1943–1944*. Oxford: Oxford University Press.

Klass, M. 1978. *From field to factory: Community structure and industrialization in West Bengal*. Philadelphia: Institute for the Study of Human Issues.

Kolenda, P. 1978. *Caste in contemporary India*. Menlo Park: The Benjamin/ Cummings Publishing Co.

Maddison, A. 1971. *Class structure and economic growth. India and Pakistan since the Moghuls*. New York: W.W. Norton and Co.

Morris, M. 1960. Caste and the evolution of the industrial workforce. *Proceedings of the American Philosophical Society* 104: 124–133.

Mukerjee, R. 1937. Caste and social change in India. *American Journal of Sociology* 43: 377–390.

Platteau, J. 1995. An Indian model of aristocratic patronage. *Oxford Economic Papers* 47: 636–662.

Rudner, D. 1994. *Caste and capitalism in colonial India: The Nattukottai Chettiars*. Berkeley: University of California Press.

Srinivas, M. 1962. *Caste in modern India and other essays*. London: Asia Publishing House.

Catallactics

Murray N. Rothbard

Keywords

Catallactics; Condillac, E. B. de; Crusoe economics; Exchange; Macleod, H. D.; Mises, L. E. von; Perry, A. L.; Political economy; Praxeology; Property; Schumpeter, J. A.; Smith, A.; Subjective theory of value; Wealth

JEL Classifications

B1

The term, meaning 'the science of exchanges', was proposed as a replacement for the name 'political economy' by the Rev. Richard Whately in his 1831 Drummond Lectures at Oxford on political economy (Whately 1831). As the leader of the group of embattled religious and economic liberals at Oriel College, Oxford, during the 1820s, Whately, a distinguished logician, had become tutor and lifelong friend of the economist Nassau W. Senior. In his Drummond Lectures, Whately was concerned to refute the dominant Oxford view that political economy, being concerned with wealth, was materialistic and opposed to Christianity. In focusing on exchanges, Whately denounced Adam Smith's definition of the scope of political economy as the science of wealth.

Whately defined man as 'an animal that makes exchanges', pointing out that even the animals nearest to rationality have not 'to all appearance, the least notion of bartering, or in any way exchanging one thing for another' (Whately 1831, p. 7). Focusing on human acts of exchange rather than on the *things* being exchanged, Whately was led almost immediately to a subjective theory of value, since he saw that 'the same thing is different to different persons' (p. 8) and that differences in subjective value are the foundation of all exchanges.

In 1831 Whately was named Archbishop of Dublin, where he promptly used his influence to

create and financially support a permanent five-year Whately Chair of Political Economy at Trinity College. For the rest of his life Whately personally selected the holders of the chair; as a result, the Whately professors carried on their mentor's tradition of catallactics and subjective utility theory. In contrast to John Stuart Mill's development of economics as a science of the abstraction 'economic man', man engaged only in avaricious pursuit of wealth, the third holder of the Whately Chair, James Anthony Lawson (1817–87), developed the idea of economics as catallactics, as studying exchanging man. Lawson, holder of the chair in his twenties (1841–6), and later to become an MP and Attorney-General for Ireland, stated in his first lecture that economics views man 'in connection with his fellow-man, having reference solely to those relations which are the consequences of a particular act, to which his nature leads him, namely, the act of making exchange' (Lawson 1844, pp. 12–13). Yet, Lawson himself fell back on discussions of wealth in his second lecture, demonstrating that, in their specific exposition, the catallacticians had not yet fully emancipated themselves from the older definitions of the scope and nature of political economy (Kirzner 1960).

One pseudonymous English writer who adopted catallactics in this period was Patrick Plough, who included and explained the term in the title of his tract, *Letters on the Rudiments of a Science, called, formerly, improperly, Political Economy, recently more pertinently, Catallactics* (London, 1842).

Catallactics reached the status of a self-conscious school of thought in the writings of the zealous and indefatigable Scottish lawyer and economist Henry Dunning Macleod. Stressing value as the result of a subjective desire of the mind, Macleod furthered the emancipation of economics from material wealth by showing that immaterial goods or services are also subjects of exchange. Macleod insisted that catallactics was the only correct school of economic thought and traced back the origins of the school beyond Whately to the late 18th-century French philosopher Etienne Bonnot de Condillac. While Condillac, in his *Le commerce et le gouvernement* (1776), did not actually use the term catallactics, he defined economics as the philosophy of commerce, or the science of exchanges. Condillac also noted that value stems only from mental desires, and hence demand, for exchangeable goods, and proclaimed that men engage in exchange precisely because each man values what he gains in exchange more than what he gives up. Hence both parties to an exchange gain in value (Macleod 1863, pp. 530–5).

The catallactic school found its culmination in the United States, in Arthur Latham Perry (1830–1905), for half a century a highly influential professor of political economy at Williams College. Perry endorsed the Macleod view of the history of economic thought, the sound catallactic school descending from Condillac through Whately and Macleod. He went beyond the inconsistencies of his forerunners, however, by purging the word 'wealth' from economics altogether, and proposing the 'property' – that which can be bought and sold – be used as a term denoting valuable things not yet sold and therefore in need of an estimate of their value (Perry 1865).

While interest in the catallactic approach faded after the work of Perry, a variant appeared in the early work of Schumpeter (1908). In this manifesto for the reconstruction of economic theory, Schumpeter wished to purge economics of all concern about purposeful human motives or actions and replace it with exclusive concentration on mechanistic alterations of economic quantities. Exchanges then become 'purely formal' variations in economic quantities of goods (Schumpeter 1908, pp. 49–55, 86, 582; Machlup 1951; Kirzner 1960).

Schumpeter did, however, manage to contribute positively to the catallactic approach. Whately and his followers had strongly rejected any element of Crusoe economics, since for them economic analysis had to be confined to interpersonal exchange. In Schumpeter's formalistic approach, actions of Crusoe could alter the placement of quantities of economic goods and therefore could be considered 'exchanges'.

It remained for Ludwig von Mises (1949) to bring back the term catallactics in his treatise on economics, and to broaden it by embedding its analysis of the market, or the science of

exchanges, in the wider discipline of 'praxeology', the science of human action. Crusoe economics then becomes vindicated in the broader sense of analysing Crusoe's actions and his use of resources to achieve his values and goals, as well as in the sense of exchanging his present state for a more satisfying one.

See Also

► Macleod, Henry Dunning (1821–1902)

Bibliography

de Condillac, E.B. 1776. Le commerce et le gouvernement considérés relativement l'un à l'autre. In *Oeuvres philosophiques de Condillac,* ed. George LeRoy. Vol. 2. Paris: Presses Universitaires de France, 1947–51.

Kirzner, I.M. 1960. *The economic point of view: An essay in the history of economic thought*. Princeton: Van Nostrand.

Lawson, J.A. 1844. *Five lectures on political economy.* Delivered before the University of Dublin, 1843. London and Dublin.

Machlup, F. 1951. Schumpeter's economic methodology. *Review of Economics and Statistics* 33: 145–151.

Macleod, H.D. 1863. *A dictionary of political economy.* Vol. 1. London.

Perry, A.L. 1865. *Political economy.* 21st ed. New York: Scribners, 1892.

Plough, P. 1842. *Letters on the rudiments of a science, called, formerly, improperly, political economy, recently more pertinently.* London: Catallactics.

Schumpeter, J.A. 1908. *Das Wesen und der Hauptinhalt der theoretischen Nationalökonomie*. Leipzig: Duncker & Humblot.

von Mises, L. 1949. *Human action: A treatise on economics*. New Haven: Yale University Press.

Whately, R. 1831. *Introductory lectures on political economy.* 2nd ed. London, 1832.

Catastrophe Theory

Y. Balasko

The theory of general equilibrium defines equilibrium prices p as the solutions in the commodity space of the vector equation defined by equality of supply and demand, namely $z(p) = 0$, where z denotes aggregate excess demand. This formulation leads to a purely mathematical problem, namely the study of the properties of the solutions of the equation $z(p) = 0$. The first problem to come into the picture is that of existence. Its positive solution leads to new issues such as the determinateness of the solutions or their number. The fact that these problems cannot be solved uniformly with exactly the same answer for every economy necessitates the introduction of suitable parameters in terms of which the properties of the solutions of the equilibrium equation can be properly described. Let ω denote this parameter chosen in some suitable vector space Ω. This means that the aggregate demand function z can be viewed as depending on ω 0 Ω which we now denote by $z(., \omega)$ and the goal of equilibrium theory becomes one of relating the properties of the solutions to $z(p, \omega) = 0$ with the parameter ω. In practice, one chooses for ω the initial endowments of every consumer, the equilibrium model simply describing a pure exchange economy.

This way of handling problems by parameterizing them had been introduced by Poincaré, who called it the continuation method. It has also been extensively used by engineers dealing with applied issues involving solving equations dependent on parameters. The topic popularized by Thom under the name catastrophe theory consists simply in combining Poincaré's continuity method with the tools of singularity theory. As a first approximation, a singularity is just another word for a multiple root of the equation $z(p, \omega) = 0$ where the unknown is the vector p. One easily sees, at least intuitively, that multiple roots, and especially double roots, correspond to borderline cases associated with changes in the number of solutions, the standard picture being that solutions appear or disappear in pairs at these double roots. Clearly enough, this may entail discontinuous behaviour of the equilibrium solution despite the fact every other feature of the model is continuous or even smooth. Catastrophe theory has often been unduly identified to this discontinuity property.

In a pure exchange economy consisting of l commodities and m consumers, the parameter

space Ω, namely the set of initial endowments, can be identified to $(R^1)^m$. Prices can conveniently be normalized, for example, with the help of the numeraire convention, so that the price space can be identified to $S = \mathbb{R}^{l-1}_{++}$ Then, the problem is to describe the set E of solutions (p, ω) to $z(p, \omega) = 0$, i.e., $E = \{(p, \omega) \in S \times \boldsymbol{\Omega} / z(p, \omega) = 0\}$ (global approach), and the solutions $(p, \omega) \in E$ when ω varies (local approach). The main results are the following ones:

1. Under smoothness assumptions for preferences, E is a smooth submanifold of $S \times \boldsymbol{\Omega}$ diffeomorphic to $\boldsymbol{\Omega}$. Furthermore, the natural projection $\pi : E \to \boldsymbol{\Omega}$ defined by the formula $(p, \omega) \to \omega$ is proper (and smooth).
2. The set Σ consisting of $\omega \in \boldsymbol{\Omega}$ for which the equilibrium equation possesses a multiple root is closed with Lebesgue measure zero.
3. Let P be the set of Pareto optima. This subset does not intersect Σ. Furthermore, there is uniqueness of equilibrium when ω describes the connected component containing the set of Pareto optima P in the complement of the set Σ in $\boldsymbol{\Omega}$.

This latter result implies that, for an economy where the trade vector remains small to some extent, equilibrium is unique and depends smoothly on the parameters defining the economy. On the other hand, when this trade vector is large, the economy is likely to have multiple equilibria so that, when the parameter vector ω varies, 'catastrophic' changes of the equilibrium prices and allocations are susceptible of being observed.

These relationships between the properties of equilibria and their number are special cases of a far more general property of the general equilibrium model. We state it as follows. Let $N(\omega)$ denotes the number of solutions of the equilibrium equation $z(p, \omega) = 0$, with $p \in S$. Then, assume that N is given (i.e., the number of solutions of the equilibrium equation is known for every $\omega \in \boldsymbol{\Omega}$). Furthermore, assume there exists an economy ω with at least two equilibria, i.e., $N(\omega) \geq 2$. Then, there is enough information to determine all the equilibrium prices associated with every economy $\omega \in \boldsymbol{\Omega}$. In other words, the economic model possesses the quite remarkable property that knowing the number of solutions suffices to determine the precise value of these solutions (provided there is an economy with multiple equilibria). If there is uniqueness of equilibrium, the above statement does not hold true any more. In that case, one finds that this unique equilibrium price vector is constant, i.e., does not depend on the economy $\omega \in \boldsymbol{\Omega}$.

See Also

- ▶ Global Analysis in Economic Theory
- ▶ Regular Economies

Catastrophic Risk

Richard A. Posner

Abstract

Catastrophic risks are defined here as events of low or unknown probability that if they occur inflict enormous losses often having a large non-monetary component. The Indian Ocean tsunami of 2004 is at the lower level of the catastrophic-risk scale of destruction; examples from higher levels including large asteroid strikes, pandemics and global warming. The challenge is to modify the principles of cost–benefit analysis to deal with serious problems caused by uncertainty (as distinct from risk), nonlinearity in value-of-life estimates, the need to project social discount rates into the distant future, and the difficulty of devising suitable policy instruments.

Keywords

Catastrophic risk; Climate change, economics of; Cost–benefit analysis; Discount rate; Carbon emission tax; Global warming; Inverse cost–benefit analysis; Kyoto Protocol; Research and development (R&D); Risk; Uncertainty; Value of life

JEL Classifications
D81

The Indian Ocean tsunami of December 2004 and, less than a year later, the flooding of New Orleans as a result of Hurricane Katrina focused attention on a type of disaster to which policymakers pay too little attention – a disaster that has a low or unknown probability of occurring but that, if it does occur, creates enormous losses. Great as were the death toll, the physical and emotional suffering of survivors, and property damage caused by the tsunami, and the even greater property damage caused by the flooding of New Orleans, even greater losses could be inflicted by other disasters of low (but not negligible) or unknown probability. The asteroid that exploded above Siberia in 1908 with the force of a hydrogen bomb might have killed millions of people had it exploded above a major city. Yet that asteroid was only about 200 feet in diameter, and a much larger one (among the thousands of dangerously large asteroids in orbits that intersect the earth's orbit) could strike the earth and, wherever it struck, cause the total extinction of the human race through a combination of shock waves, fire, tsunamis, and blockage of sunlight. Other catastrophic risks include, besides earthquakes such as the one that caused the 2004 tsunami, natural epidemics (the 1918–19 Spanish influenza epidemic killed between 20 million and 40 million people), nuclear or biological attacks by terrorists, certain types of lab accident (one discussed later in this article), and abrupt global warming. The probability of catastrophes resulting, whether or not intentionally, from human activity appears to be increasing because of the rapidity and direction of technological advances.

The Economic Approach to Catastrophe

It is generally believed that the prediction, assessment, prevention, and mitigation of catastrophes is the province of science. However, economic analysis has an important role to play, as well. Able scientists can commit analytical errors when discussing policy that economists would easily avoid. Thus, Barry Bloom, dean of the Harvard School of Public Health, has criticized the editors of leading scientific journals for having taken the position that 'an editor may conclude that the potential harm of publication outweighs the potential societal benefits' (Bloom 2003, pp. 48, 51). (The specific reference is to publications from which terrorists could learn how to create lethal bioweapons.) Bloom calls this 'a chilling example of the impact of terrorism on the freedom of inquiry and dissemination of knowledge that today challenges every research university' (Bloom 2003, p. 51). The implication – that freedom of scientific research should enjoy absolute priority over every other social value – neglects the need to weigh costs and benefits in order to determine the best balance between public safety and scientific progress.

To illustrate the economic approach to catastrophe, suppose that a tsunami as destructive as the Indian Ocean tsunami occurs on average once a century and kills 250,000 people. That is an average of 2500 deaths per year. Even without attempting a sophisticated estimate of the value of life to the people exposed to the risk, one can say with some confidence that, if an annual death toll of 2500 could be substantially reduced at moderate cost, the investment would be worthwhile. A combination of educating the residents of low-lying coastal areas about the warning signs of a tsunami (tremors and a sudden recession in the ocean), establishing a warning system involving emergency broadcasts, telephoned warnings, and air-raid- type sirens, and improving emergency response systems would have saved many of the people killed by the Indian Ocean tsunami, probably at a total cost below any reasonable estimate of the average losses that can be expected from tsunamis. Relocating people away from coasts would be even more efficacious, but, except in the most vulnerable areas or in areas in which residential or commercial uses have only marginal value, the costs would probably exceed the benefits. For annual costs of protection must be matched with annual, not total, expected costs of tsunamis.

As another example, consider the question of optimal precautions against the type of flood that

inundated New Orleans. In 1998 it was estimated that it would cost $14 billion to prevent such a flood; the estimated 'economic' cost (which ignores the loss of life and physical and emotional suffering) of the recent flood is $100 billion to $200 billion; and the Corps of Engineers estimated the annual probability of such a flood at 1 in 300. If we take the lower cost and assume that the $14 billion investment would eliminate the probability of a flood within 30 years, a period in which the probability of a flood if the measures were not taken would be a shade under ten per cent, yielding an expected benefit from the flood-control measures of $10 billion, the measures would flunk a cost–benefit test. Note that the calculation does not include discounting future benefits to present value; the reason is that the benefits are likely to grow – a flood that occurred 30 years hence would be likely to do more damage because property values would increase.

Value of Life Estimates

What might tip the balance in favour of the flood-control measures would be monetizing the expected loss of life and other human suffering. There is now a substantial economic literature inferring the value of life from the costs people are willing to incur to avoid small risks of death; if from behaviour toward risk one infers that a person would pay $70 to avoid a 1 in 100,000 risk of death, his value of life would be estimated at $7 million ($70/.00001), which is in fact the median estimate of the value of life of an American (Viscusi and Aldy 2003, pp. 5, 18, 63). The value of this transformation is simply that, once a risk is calculated, its expected cost is instantly derived simply by multiplying the risk by the value of life.

But there is significant nonlinearity to be considered at both ends of the risk spectrum. At the high end, if one is asked what he would demand to play one round of Russian roulette, the typical answer will be a good deal more than 1/6 of $7 million. At the low-probability end of the risk spectrum, there is a tendency to write the cost of the risk down to or near zero (see, for example, Kunreuther and Pauly 2004; Viscusi 1997). In other words, the studies from which the $7 million figure is derived may not be robust with respect to risks of death either much larger or much smaller than the 1 in 10,000 to 1 in 100,000 range of most of the studies – and we do not know what the risk of death from a tsunami was to the people killed, though it was probably towards the low end of the range.

Even if we disregard this issue, because value of life is positively correlated with income, the $7 million figure cannot be used to estimate the value of life of the people killed by the Indian Ocean tsunami, or at least most of them (and perhaps likewise the people killed in the New Orleans flood, most of whom were poor). Additional complications arise from the fact that the deaths were only a part of the cost inflicted by the disaster – the injuries, the suffering, and the property damage that also resulted from the tsunami have to be estimated along with the efficacy and expense of precautionary measures that would have been feasible. The risks of smaller but still destructive tsunamis that such measures might protect against must also be factored in; nor is the 'once a century' risk estimate much better than a guess. Nevertheless, it seems apparent that the total cost of the tsunami was high enough to indicate that precautionary measures would have been cost-justified.

The tsunami, unlike the New Orleans flood, could not have been prevented. The only possible precautionary measures would have been either a warning system to enable prompt evacuation or permanently relocating population away from the coastline. Similar measures would have been possible alternatives to preventive measures for New Orleans as well, especially a system for prompt evacuation; but such a system would not have prevented either property damage or massive if temporary population relocation, both of which were huge costs of the flood.

The Political Economy of Catastrophe Prevention and Response

Since precautionary measures of some kind taken in anticipation of a tsunami on the scale that

occurred would clearly have been cost-justified, why were they not taken? Tsunamis are a common consequence of earthquakes, which themselves are common; and tsunamis can have other causes besides earthquakes – a major asteroid strike in an ocean would create a tsunami that would dwarf the Indian Ocean one. The answer, or answers, may be economic in character.

First, although a once-in-a-century event is as likely to occur at the beginning of the century as at any other time, it is much less likely to occur some time in the first decade of the century than some time in the last nine decades of the century. (The point is simply that the probability is greater the longer the interval being considered: one is more likely to catch a cold in the next year than in the next 48 hours.) Politicians with limited terms of office and thus foreshortened political horizons are likely to discount low-risk disaster possibilities steeply because the risk of damage to their careers from failing to take precautionary measures is truncated.

Second, to the extent that effective precautions require governmental action, the fact that government is a centralized system of control makes it difficult for officials to respond to the full spectrum of possible risks against which cost-justified measures might be taken. Given the variety of matters to which they must attend, officials are likely to have a high threshold of attention below which risks are simply ignored. The US government, preoccupied with terrorist threats, paid insufficient attention to the risk of a disastrous flood of New Orleans, though the risk was understood to be significant.

Third, where risks are regional or global rather than local, many national governments, especially in the poorer and smaller countries, may drag their heels in the hope of taking a free ride on the larger and richer countries. Knowing this, the latter countries may be reluctant to take precautionary measures and by doing so reward and thus encourage free riding. Again, there is a US parallel: state and local government may stint on devoting resources to emergency response, expecting aid from other state and local governments and the federal government.

Fourth, countries are poor often because of weak, inefficient, or corrupt government, characteristics that may disable poor nations from taking cost-justified precautions. Again there is a US parallel: Louisiana is a poor state and New Orleans, which has a very large poor population, has a reputation for having an inefficient and even corrupt government.

And fifth, the positive correlation of per capita income with value of life suggests that it is quite rational for even a well-governed poor country to devote proportionately fewer resources to averting calamities than rich countries do. This would also be true of a poor state or city of the United States.

The failure to act in accordance with cost–benefit principles is dominant characteristic of public policy towards catastrophic risk. An example is the asteroid menace, which is analytically similar to the menace of tsunamis. The National Aeronautics and Space Administration, with an annual budget of more than $10 billion, spends only $4 million a year on mapping dangerously close large asteroids, and at that rate may not complete the task for another decade, even though such mapping is the key to an asteroid defence because it may provide many years of advance warning. Deflecting an asteroid from its orbit when it is still hundreds of millions of miles away from hitting the earth appears to be a feasible undertaking. Although asteroid strikes are less frequent than tsunamis, there have been enough of them to enable the annual probabilities of various magnitudes of such strikes to be estimated, and from these estimates an expected cost of asteroid damage can be calculated. As in the case of tsunamis, if there are measures, beyond those being taken already, that can reduce the expected cost of asteroid damage at a lower cost, thus yielding a net benefit, the measures should be taken, or at least seriously considered.

Cost–Benefit Analysis Under Uncertainty

Often it is not possible to estimate the probability or magnitude of a possible catastrophe; the situation is one of uncertainty rather than of risk; how

then can cost–benefit analysis, or other techniques of economic analysis, help us in devising responses to such a possibility? The probability of bioterrorism or nuclear terrorism, for example, cannot be quantified; nevertheless, there is rough sense of the range of possible losses that such terrorism would inflict – a range that has no upper limit short of the extinction of the human race – and from this it can be inferred that, even if the probability of such a terrorist attack is small, the expected cost – the product of the probability of the attack and of the consequences if the attack occurs – probably is quite high.

An example of how economic analysis can produce insights even when catastrophic risks are non-quantifiable involves the Relativistic Heavy Ion Collider (RHIC) that went into operation at Brookhaven National Laboratory in Long Island in 2000. As explained by the distinguished English physicist Sir Martin Rees, the collisions in RHIC might conceivably produce a shower of quarks that would 'reassemble themselves into a very compressed object called a strangelet. . . . A strangelet could, by contagion, convert anything else it encountered into a strange new form of matter. . .. A hypothetical strangelet disaster could transform the entire planet Earth into an inert hyperdense sphere about one hundred metres across' (Rees 2003, pp. 120–1). Rees considers this 'hypothetical scenario' exceedingly unlikely, yet points out that even an annual probability of 1 in 500 million is not wholly negligible when the result, should the improbable materialize, would be so total a disaster.

Concern with such a possibility led John Marburger, the director of the Brookhaven National Laboratory, to commission a risk assessment by a committee of distinguished physicists before authorizing RHIC to begin operating. The committee concluded that the risk of a strangelet disaster was negligible. No cost–benefit analysis of RHIC was conducted, with or without including the risk of a strangelet disaster on the cost side. RHIC cost $600 million to build, and its annual operating costs were expected to be $130 million. No attempt was made to monetize the benefits that the experiments conducted in it were expected to yield; because the experiments are designed to satisfy scientific curiosity rather than to create knowledge that is likely to lead to the invention of useful products, estimation of the benefits is impossible. They may be slight.

The probability of a strangelet disaster in the course of RHIC's planned ten-year life cannot actually be quantified, though there have been attempts. One team of physicists estimated the probability of a strangelet disaster as no more than 1 in 50 million. The official risk-assessment team offered a series of upper-bound estimates, including a 1 in 500,000 probability of a strangelet disaster over the ten-year period, which is 100 times greater than the other's team's estimate. These really are wild, as well as wildly divergent, guesses. Still another uncertainty is what dollar figure to place on the destruction of the earth and all its human and other inhabitants, given the nonlinearity of value of life estimates. Yet, given these uncertainties, the fact that the benefits of RHIC may be quite small suggests that the possibility, remote as it may seem, of a strangelet disaster would weigh heavily, in an economic analysis, against the project. There are more than six billion people on Earth – not to mention unborn future generations – and if their average value of life is estimated at a modest $1 million, the cost of extinction would be $6 quadrillion, and a 1 in 100 million annual risk of a strangelet disaster would yield an annual expected extinction cost of $60 million for ten years to add to the $130 million in annual operating costs and the initial investment of $600 million – roughly a one-third increase in total cost. This could well be decisive against the project, given its entirely conjectural benefits.

Global Warming: Risk And Response

Another, more familiar, example of the difficulty of quantifying catastrophic risk is the problem of global warming. The Kyoto Protocol, which came into effect by its terms when Russia signed it although the United States has not done so, requires the signatory nations to reduce their carbon dioxide emissions to a level seven to ten per cent below what they were in the late 1990s, but

exempts developing countries, such as China, a large and growing emitter, and Brazil, which is destroying large reaches of the Amazon rainforest, much of it by burning. The effect of carbon dioxide emissions on the atmospheric concentration of the gas is cumulative, because carbon dioxide leaves the atmosphere (by being absorbed into the oceans) at a much lower rate than it enters it, and therefore the concentration will continue to grow even if the annual rate of emission is cut down substantially. Between this phenomenon and the exemptions, there is a widespread belief that the Kyoto Protocol will have only a slight effect in arresting global warming; yet the tax or other regulatory measures required to reduce emissions below their level of six years ago will be very costly.

The Protocol's supporters generally are content to slow the rate of global warming by encouraging – by means of heavy taxes (for example, on gasoline or coal) or other measures (such as quotas) that will make fossil fuels more expensive to consumers –conservation measures such as driving less or driving more fuel- efficient cars that will reduce the consumption of these fuels. But from an economic standpoint that is probably either too much or too little. It is too much if, as most scientists believe, global warming will continue to be a gradual process, producing really serious effects – the destruction of tropical agriculture, the spread of tropical diseases such as malaria to currently temperate zones, dramatic increases in violent storm activity (increased atmospheric temperatures, by increasing the amount of water vapour in the atmosphere, increase precipitation), and a rise in sea levels (eventually to the point of inundating most coastal cities) – only toward the end of the 21st century. By that time science, without prodding by governments, is likely to have developed economical 'clean' substitutes for fossil fuels (there already is a clean substitute – nuclear power) and even economical technologies for either preventing carbon dioxide from being emitted into the atmosphere by the burning of fossil fuels, or removing it from the atmosphere.

But the Protocol is too little and too late, as a response to the costs of global warming, if the focus is changed from gradual to abrupt global warming. At various times in the Earth's history, drastic temperature changes have occurred in the course of just a few years. During the Younger Dryas epoch of about 11,000 years ago, shortly after the end of the last ice age, global temperatures soared by about 14 degrees Fahrenheit in the course of a decade. Because the earth was still cool from the ice age, the effect of the increased warmth on the human population was positive. But a similar increase in a modern decade would have devastating effects on agriculture and on coastal cities, and might even cause a shift in the Gulf Stream that would result in giving all of Europe a Siberian climate.

Because of the enormous complexity of the forces that determine climate, and the historically unprecedented magnitude of human effects on the concentration of greenhouse gases, the possibility that continued growth in that concentration could precipitate – and within the near rather than the distant future – a sudden warming similar to that of the Younger Dryas cannot be excluded. Indeed, no probability, high or low, can be assigned to such a catastrophe. But it may be significant that, while dissent continues, many climate scientists are now predicting dramatic effects from global warming within the next 20 to 40 years, rather than just by the end of the century (Lempinen 2005). It may be prudent, therefore, to try to stimulate an increase in the rate at which economical substitutes for fossil fuels, and technology both for limiting the emission of carbon dioxide by those fuels when they are burned in internal-combustion engines or electrical generating plants, and for removing carbon dioxide from the atmosphere, are developed. This can be done by stiff taxes on carbon dioxide emissions. Such taxes give the energy industries, along with customers of theirs such as airlines and manufacturers of motor vehicles, a strong incentive to finance R&D designed to create economical clean substitutes for such fuels and devices to 'trap' emissions at the source before they enter the atmosphere. Given the technological predominance of the United States, it is important that these taxes be imposed on US firms, which they would be if the United States ratified the Kyoto Protocol.

One advantage of the technology-forcing tax approach over public subsidies for R&D is that the government would not be in the business of picking winners – the affected industries would decide what R&D to support – and another is that the brunt of the taxes could be partly offset by reducing other taxes, since emission taxes would raise revenue as well as inducing greater R&D expenditures.

It might seem that subsidies would be necessary for technologies that would have no market, such as technologies for removing carbon dioxide from the atmosphere. There would be no private demand for such technologies because, in contrast to ones that reduce emissions, technologies that remove already emitted carbon dioxide from the atmosphere would not reduce any emitter's tax burden. But this problem is easily solved by making the tax a tax on *net* emissions. Then an electrical generating plant or other emitter could reduce its tax burden by removing carbon dioxide from the atmosphere as well as by reducing its own emissions of carbon dioxide into the atmosphere.

It might seem that, because the demand for conventional fuel sources is inelastic in the short run, the imposition of stiff taxes or quotas required by the Kyoto Protocol would have little effect on the level of emissions. But the significance of the taxes, which actually depends on the inelasticity of demand, is that it would create both pressures and resources for finding a technological fix that would counter the cumulative effect of emissions on the atmospheric concentration of carbon dioxide by driving annual emissions to zero or even below.

Global Warming: the Discounting Problem

A further advantage of focusing on the risk of abrupt rather than gradual global warming is that it allows the vexing problem of discount rate to be elided. The problem is acute when concern focuses on gradual global warming. Suppose that a $10 billion expenditure on capping emissions today would have no effect on human welfare during this century but, by slowing global warming, would produce a savings in social costs of $100 billion in 2100. At a discount rate of three per cent, the present value of $100 billion a century from now is only $5 billion. That would make the expenditure of $10 billion today seem a very poor investment. (For the sake of simplicity, benefits that are expected to accrue after 2100 are ignored in this analysis.) The same amount of money invested in financial instruments could be expected to grow to $192 billion by 2100, on the assumption of a three per cent real interest rate for the next 100 years (though in fact interest rates cannot be forecast over such a long period). If the fund were then disbursed to the victims of global warming, they would be better off than if the $100 billion cost of global warming assumed to be incurred in that year had been averted. Less conservative investments, moreover, would yield larger expected returns – ten per cent or more rather than three per cent.

But it is not a real alternative to spending $10 billion now to invest it in a fund for future victims of global warming. No such fund will be created, and so they will not be compensated. In circumstances such as this, discounting future to present values is not a method of helping people to decide how to manage their affairs in the way most conducive to maximizing their welfare. Rather, it is a method of maximizing global wealth without regard to its distribution among persons. In the case of gradual global warming, the victims are likely to be concentrated in poor countries, so that basing policy on the discounted costs of global warming would further immiserate the future inhabitants of those countries by increasing the authorized level of emissions harmful to them.

A discount rate based on market interest rates tends to obliterate the interests of remote future generations. The implications are drastic. 'At a discount rate of five per cent, one death next year counts for more than a billion deaths in 500 years. On this view, catastrophes in the further future can now be regarded as morally trivial' (Parfit 1984, p. 357). (What right would the Romans have had to regard our lives as worthless in deciding whether to conduct dangerous experiments?) The trade-off is only slightly less extreme if one substitutes 100 years for 500. At a five per cent discount rate, the present value of one dollar to be received in 100 years is only three-quarters of a cent – and if

for money we substitute lives, then to save one life this year we should be willing to sacrifice almost 150 lives a century hence.

And yet not to discount future costs at all would be absurd, certainly as a practical political matter. For then the present value of benefits conferred on our remote descendants would approach infinity. Measures taken today to arrest global warming would confer benefits not only in 2100 but in every subsequent year, perhaps for millions of years. The present value of $100 billion received every year for a million years at a discount rate of zero per cent is $100 quadrillion.

But the vexing problem of how much weight to give to the welfare of remote future generations can be finessed, at least to some extent, if not solved. A discounted present value can be equated to an undiscounted present value simply by shortening the time horizon for the consideration of costs and benefits. For example, the present value of an infinite stream of costs discounted at four per cent a year is equal to the undiscounted sum of those costs for 25 years, while the present value of an infinite stream of costs discounted at one per cent a year is equal to the undiscounted sum of those costs for 100 years. The formula for the present value of one dollar per year forever is $1/r, where r is the discount rate. So if r is four per cent, the present value is $25, and this is equal to an undiscounted stream of one dollar per year for 25 years. If r is one per cent, the undiscounted equivalent is 100 years.

One way to argue for the four per cent rate (that is, for truncating our concern for future welfare at 25 years) is to say that we're willing to weight the welfare of the next generation as heavily as our own welfare but that's the extent of our regard for the future. One way to argue for the one per cent rate is to say that we are willing to give equal weight to the welfare of everyone living in this century, which will include us, our children, and our grandchildren, but beyond that we don't care. Looking at future welfare in this way, we may be inclined towards the lower rate, which would have dramatic implications for willingness to invest today in limiting global warming. The lower rate could even be regarded as a ceiling. Most people have some regard for human welfare, or at least the survival of some human civilization, in future centuries. We are grateful that the Romans didn't exterminate the human race in chagrin at the impending collapse of their empire.

Another way to bring future consequences into focus without conventional discounting is by aggregating risks over time rather than expressing them in annualized terms. If we are concerned about what may happen over the next century, then instead of asking what the annual probability of a collision with a ten-kilometre- wide asteroid is, we might ask what the probability is that such a collision will occur within the next 100 years. An annual probability of 1 in 75 million translates into a century probability of roughly 1 in 750,000. That may be high enough – in view of the consequences if the risk materializes – to justify spending several hundred million, perhaps even several billion, dollars to avert it.

Inverse Cost–Benefit Analysis

A helpful approach to cost–benefit analysis under conditions of extreme uncertainty is what can be called 'inverse cost–benefit analysis' (Posner 2004, pp. 176–84). Analogous to extracting probability estimates from insurance premiums, it involves dividing what the government is spending to prevent a particular catastrophic risk from materializing by what the social cost of the catastrophe would be if it did materialize. The result is an approximation to the implied probability of the catastrophe. Expected cost is the product of probability and consequence (loss): $C = PL$. If P and L are known, C can be calculated. If instead C and L are known, P can be calculated: if $1 billion ($C$) is being spent to avert a disaster that if it occurs will impose a loss (L) of $100 billion, then $P = C/L = .01$.

If P so calculated diverges sharply from independent estimates of it, this is a clue that society may be spending too much or too little on avoiding L. It is just a clue, because of the distinction between marginal and total costs and benefits. The optimal expenditure on a measure is the expenditure that equates marginal cost to marginal benefit. Suppose we happen to know that P is not .01 but .1, so that the expected cost of the

catastrophe is not $1 billion but $10 billion. It doesn't follow that we should be spending $10 billion, or indeed anything more than $1 billion, to avert the catastrophe. Perhaps spending just $1 billion would reduce the expected cost of catastrophe from $10 billion all the way down to $500 million and no further expenditure would bring about a further reduction, or at least a cost-justified reduction. For example, if spending another $1 billion would reduce the expected cost from $500 million to zero, that would be a bad investment, at least if risk aversion is ignored.

The federal government is spending about $2 billion a year to prevent a bioterrorist attack (increased to $2.5 billion for 2005 under the rubric of 'Project BioShield') (Office of Management and Budget 2003, pp. 37–8; US Department of Homeland Security 2004). The goal is to protect Americans, so in assessing the benefits of this expenditure casualties in other countries can be ignored. Suppose the most destructive biological attack that seems reasonably possible on the basis of what little we now know about terrorist intentions and capabilities would kill 100 million Americans. We know that value-of-life estimates may have to be radically discounted when the probability of death is exceedingly slight. But there is no convincing reason for supposing the probability of such an attack less than, say, 1 in 100,000; and the value of life that is derived by dividing the cost that Americans will incur to avoid a risk of death of that magnitude by the risk is about $7 million. Then, if the attack occurred, the total costs would be $700 trillion – and that is actually too low an estimate because the death of a third of the population would have all sorts of collateral consequences, mainly negative. Let us, still conservatively however, refigure the total costs as $1 quadrillion. The result of dividing the money being spent to prevent such an attack, $2 billion, by $1 quadrillion is 1/500,000. Is there only a 1 in 500,000 probability of a bioterrorist attack of that magnitude in the next year? One doesn't know, but the figure seems too low.

It doesn't follow that $2 billion a year is too little to be spending to prevent a bioterrorist attack; one must not forget the distinction between total and marginal costs. Suppose that the $2 billion expenditure reduces the probability of such an attack from .01 to .0001. The expected cost of the attack would still be very high – $1 quadrillion multiplied by .0001 is $100 billion – but spending more than $2 billion might not reduce the residual probability of .0001 at all. For there might be no feasible further measures to take to combat bioterrorism, especially when we remember that increasing the number of people involved in defending against bioterrorism, including not only scientific and technical personnel but also security guards in laboratories where lethal pathogens are stored, also increases the number of people capable, alone or in conjunction with others, of mounting biological attacks. But there *are* other response measures that should be considered seriously. And one must also bear in mind that expenditures on combating bioterrorism do more than prevent mega-attacks; the lesser attacks, which would still be very costly both singly and cumulatively, would also be prevented.

Costs, moreover, tend to be inverse to time. It would cost a great deal more to build an asteroid defence in one year than in ten years because of the extra costs that would be required for a hasty reallocation of the required labour and capital from the current projects in which they are employed. And so would other crash efforts to prevent catastrophes. Placing a lid on current expenditures would have the incidental benefit of enabling additional expenditures to be deferred to a time when, because more will be known about both the catastrophic risks and the optimal responses to them, considerable cost savings may be possible. The case for such a ceiling derives from comparing marginal benefits to marginal costs; the latter may be sharply increasing in the short run.

See Also

- ▶ Climate Change, Economics of
- ▶ Cost–Benefit Analysis
- ▶ Environmental Economics
- ▶ Risk
- ▶ Social Discount Rate
- ▶ Value of Life

Bibliography

Bloom, B. 2003. Bioterrorism and the university: The threats to security – and to openness. *Harvard Magazine* 106(2): 48–52.

Kunreuther, H., and M. Pauly. 2004. Neglecting disaster: Why don't people insure against large losses? *Journal of Risk and Uncertainty* 28: 5–21.

Lempinen, E. 2005. Scientists on AAAS panel warn that ocean warming is having dramatic impact. AAAS News Release, 17 February. Online. Available at http://www.aaas.org/new/releases/2005/0217warmingwarning.shtml. Accessed 21 Dec 2005.

Office of Management and Budget. 2003. *2003 Report to Congress on Combating Terrorism*. Washington, DC: Executive Office of the President. Online. Available at http://www.whitehouse.gov/omb/inforeg/2003_combat_terr.pdf. Accessed 22 Dec 2005

Parfit, D. 1984. *Reasons and persons*. Oxford: Oxford University Press.

Posner, R. 2004. *Catastrophe: Risk and response*. New York: Oxford University Press.

Rees, M. 2003. *Our final hour: A scientist's warning; how terror, error, and environmental disaster threaten humankind's future in this century – on earth and beyond*. New York: Basic Books.

US Department of Homeland Security. 2004. Fact Sheet: Department of Homeland Security Appropriations Act of 2005. 18 October. Online. Available at http://www.dhs.gov/dhspublic/interapp/press_release/press_release_0541.xml. Accessed 22 Dec 2005.

Viscusi, W. 1997. Economic and psychological aspects of valuing risk reduction. In *Determining the value of non-marketed goods: Economic, psychological, and policy relevant aspects of contingent valuation methods*, ed. R. Kopp, W. Pommerehne, and N. Schwarz. Boston: Kluwer.

Viscusi, W., and Jo Aldy. 2003. The value of a statistical life: A critical review of market estimates throughout the world. *Journal of Risk and Uncertainty* 27: 5–76.

Catchings, Waddill (1879–1967)

Robert W. Dimand

Keywords

Business cycle; Catchings, W; Federal Reserve System; Foster, W; Harrod–Domar growth theory; Keynesianism; Short run and long run; Monetarism; Pollak Foundation for Economic Research

JEL Classification

B31

An investment banker and heterodox monetary economist, Waddill Catchings was born in Sewanee, Tennessee, on 6 September 1879, and died in Pompano Beach, Florida, on 31 December 1967. He graduated from Harvard College in 1901 and Harvard Law School in 1904. Joining the New York City law firm Sullivan & Cromwell on a salary of ten dollars a week, Catchings proved skilful in managing the affairs of companies that went into receivership during the financial panic of 1907, and became president of three ironworks. During the First World War, Catchings worked in the export department of J. P. Morgan & Company, then the US purchasing agent for the British and French governments. A Harvard classmate of Arthur Sachs, Catchings joined Goldman, Sachs & Company in 1918 as partner in charge of underwriting, helping to organize General Foods and National Dairy Products (later Kraft).

Catchings complained that his Harvard professors 'casually explained that their theories would hold true in the long run. But what people are interested in is the short, not the long, run. So I made up my mind that as soon as I had enough money I would set about reconciling these two phases of business – theory and practice' (quoted in his obituary in the *New York Times,* 1 January 1968). In 1920, Catchings and his Harvard classmate William Trufant Foster (a rhetoric professor and college administrator) established the Pollak Foundation for Economic Research, directed by Foster, funded by Catchings, and dedicated to promoting their belief that, in Catchings's words, 'If business is to continue zooming, production must be kept at high speed, whatever the circumstances' (*New York Times* obituary). High and growing levels of production could be maintained by high and growing levels of consumer spending, and the business cycle could be eliminated by appropriate Federal Reserve policy and by keeping public works projects in reserve for economic downturns. In addition to a syndicated newspaper column, Foster and Catchings wrote *Money* (1923), *Profits* (1925), *Business without a Buyer*

(1927), *The Road to Plenty* (1928), and *Progress and Plenty* (1930), all Pollak Foundation Studies. Gleason (1959) and Carlson (1962) consider Foster and Catchings as possible precursors of Keynesian macroeconomics and Harrod–Domar growth theory. The four per cent annual increase in currency and credit endorsed by Foster and Catchings is a possible forerunner of monetarism, but they opposed any mandating of a price level rule, preferring a goal of maintaining prosperity (Tavlas, 1976).

In December 1928, Catchings launched the Goldman Sachs Trading Corporation (GSTC), a closed-end investment trust (ten per cent owned by Goldman, Sachs & Company) which in July 1929 launched the Shenandoah Corporation, another closed-end investment trust, 40% owned by GSTC, followed in August by the Blue Ridge Corporation, with Shenandoah owning a majority of Blue Ridge's common shares. At their peak, this highly leveraged pyramid controlled $500 million of investments, but it was swept away in the stock market crash. GSTC shares, which were initially sold to the public at $104, reached $326 (thanks in part to $57 million that GSTC spent buying its own shares by March 1929, and more purchases later) before falling to $1.75. Catchings had launched Shenandoah and Blue Ridge without consulting the Sachs brothers (who were in Europe in the summer of 1929), and in May 1930 his partners forced his resignation, paying him $250,000 despite his capital account's deficit.

Catchings withdrew from the Pollak Foundation (whose endowment disappeared in the crash) to concentrate on his own finances, and moved to California. In the 1950s Catchings was a director of Chrysler, Standard Packaging, and Warner Brothers. After Foster died in 1950, Catchings collaborated with Charles F. Roos (a co-founder of the Econometric Society) on *Money, Men and Machines* (1953). Denouncing Keynesian economics, Catchings and Roos accused the Federal Reserve System of interfering with economic freedom and destabilizing the economy through roller-coaster monetary policies in futile attempts to keep higher wages from causing higher prices. Their book won the Freedoms Foundation's George Washington Honor Medal. Catchings's last books were *Do Economists Understand Business?* (1955), *Bias Against Business* (1956), and *Are We Mismanaging Money?* (1960).

See Also

- ▶ Foster, William Trufant (1879–1950)
- ▶ Monetary Cranks
- ▶ Underconsumptionism

Selected Works

1923. (With W. Foster.) *Money.* Boston: Houghton Mifflin.
1925. (With W. Foster.) *Profits.* Boston: Houghton Mifflin.
1927. (With W. Foster.) *Business without a Buyer.* Boston: Houghton Mifflin.
1928. (With W. Foster). *The Road to Plenty.* Boston: Houghton Mifflin.
1930. (With W. Foster.) *Progress and Plenty.* Boston: Houghton Mifflin.
1953. (With C. Roos.) *Money, Men and Machines.* 2nd edn. New York: Econometric Institute, 1958.
1955. *Do Economists Understand Business?* New York (privately printed).
1956. *Bias Against Business.* New York (privately printed).
1960. *Are We Mismanaging Money?* New York (privately printed).

Bibliography

Carlson, J. 1962. Foster and catchings: A mathematical reappraisal. *Journal of Political Economy* 70: 400–2.
Dorfman, J. 1959. *The economic mind in American civilization*. Volumes 4 and 5: 1918–1933. New York: Viking.
Endlich, L. 1999. *Goldman sachs: The culture of success*. New York: Knopf.
Gleason, A. 1959. Foster and catchings: A reappraisal. *Journal of Political Economy* 67: 156–72.
Tavlas, G. 1976. Some further observations on the monetary economics of Chicagoans and non-Chicagoans. *Southern Economic Journal,* 42: 685–92, with comment by J. Davis and reply by Tavlas, 45 (1979), 919–31.

Catching-Up

Stanislaw Gomulka

The search for a pattern in the observed wide variation in the cross-country growth rate of output per man hour has led to the observation that the latecomers in industrialization should, and in fact do, tend to innovate faster than does the world's 'technology frontier area' (TFA), the latter defined as the regions in which the world's best technology is employed. The reason behind this observation is the commonsense notion that in technology or organization, as well as in science, learning and imitating is typically cheaper and faster than is the original discovery and testing. The distance between the level of development of the TFA and that of a less developed country (LDC) may be taken as a measure of the backlog of technological opportunities to exploit. The larger the greater may be expected to be the economic incentive to take advantage of some of these opportunities and, other things being equal, the greater the rate of international technology transfer. The idea that there might be 'advantages of backwardness' in this sense is usually associated with the names of Thorstein Veblen and Alexander Gerschenkron. Veblen (1915) applied it to Germany vis-à-vis England; Gerschenkron (1962) updated it and extended the work to include Russia, France and Italy. A formalization of this idea by Nelson and Phelps (1966) assumed that an increase in the level of technology of an LDC is proportional to the technology gap between it and the TFA. This assumption implies that the relationship between the rate of innovation and the relative technology gap is, for any LDC and in the course of time, positive and linear. Moreover, the LDCs' innovation rate would always exceed that of the TFA but fall toward it asymptotically, the relative gap falling as a result toward a country-specific positive constant, called the 'equilibrium technology gap'. This falling of the relative technology gap between an LDC and the TFA is what is meant by (international and/or technological) catching-up.

Studies of the world pattern of productivity growth rates in the period 1950–85 have led to the important qualification of the original Veblen–Gerschenkron hypothesis, namely that for the group of highly backward LDCs, the rate of innovation tends to be lower the greater the relative technology gap. The relationship across all countries is thus of the 'hat-shaped type' (Gomulka 1971; Horvat 1974). The usual interpretation of the negative part of the Hat-shape Relationship rests on the notion of 'absorptive capacity' being the severely limiting factor in the initial phase of the catching-up. As educational standards and physical infrastructure are improved and export capabilities developed, a larger amount of foreign technology becomes profitable. Technology imports themselves also help upgrade skills and increase exports, attracting still larger technology imports, and so forth. It is this causality sequence which gives rise to the relationship's negative part. However, before absorptive capacity is developed to reach a level at which an LDC's rate of innovation is the same as that of the TFA, an LDC's relative backwardness would be increasing.

The Hat-shape Relationship may be interpreted as an international, macroeconomic equivalent of logistic or S-shaped diffusion curves observed often for individual inventions. Theoretical research has been centred on modelling the dynamics of catching-up under different channels of technology transfer, such as direct foreign investment Findlay (1976), a cost-free diffusion (Gomulka 1971), or trading conventional goods for embodied technology (Gomulka 1970). The most recent development of the theory also takes into account economic dualism and technology transfer costs. This particular theory combines international technology transfer with internal diffusion from the modern to the traditional sector, and interprets 'appropriate technology' in a dynamic context. Empirical studies indicate that embodied technology transfer is an important, perhaps the main, channel for most LDCs (Gomulka and Sylvestrowicz 1976). However, in

the postwar catching-up of the US by countries with large R and D sectors, such as Japan, West Germany and the USSR, the import of capital goods from the US has apparently played a small role, indicating that disembodied diffusion, both (virtually) cost-free and commercial, have probably played the main role. The post-1975 labour productivity slowdown in countries of the latter character may be interpreted as evidence of these highly developed countries approaching their specific equilibrium technology gaps. These equilibrium gaps, as well as innovation rates in the course of the catching-up itself, appear to be strongly influenced by cultural and systemic factors. Consequently, the process of catching-up is bringing about a state of international growth equilibrium in which the innovation rate would be common to all countries, but in which productivity and technology levels would continue to vary significantly among the world's countries.

See Also

- ▶ Backwardness
- ▶ Cumulative Causation
- ▶ Diffusion of Technology
- ▶ Periphery

References

Ames, E., and N. Rosenberg. 1963. Changing leadership and industrial growth. *Economic Journal* 73: 13–31.

Findlay, R. 1976. Relative backwardness, direct foreign investment, and the transfer of technology: A simple dynamic model. *Quarterly Journal of Economics* 92(1): 1–16.

Gerschenkron, A. 1962. *Economic backwardness in historical perspective*. Cambridge, MA: Harvard University Press.

Gomulka, S. 1970. Extensions of the 'golden rule of research' of Phelps. *Review of Economic Studies* 37: 73–93.

Gomulka, S. 1971. *Inventive activity, diffusion, and the stages of economic growth*. Aarhus: Institute of Economics.

Gomulka, S., and J.D. Sylvestrowicz. 1976. Import-led growth: Theory and estimation. In *On the measurement of factor productivities: Theoretical problems and empirical results*, ed. F.-L. Altman et al. Göttingen: Vendenhoeck and Ruprecht.

Horvat, B. 1974. Welfare and the common man in various countries. *World Development* 2(7): 29–39.

Nelson, R.R., and E.S. Phelps. 1966. Investment in humans, technological diffusion, and economic growth. *American Economic Review* 56(2): 69–75.

Veblen, T. 1915. *Imperial Germany and the industrial revolution*. London: Macmillan.

Categorical Data

A. Colin Cameron

Abstract

Categorical outcome (or discrete outcome or qualitative response) regression models are models for a discrete dependent variable recording in which of two or more categories an outcome of interest lies. For binary data (two categories) probit and logit models or semiparametric methods are used. For multinomial data (more than two categories) that are unordered, common models are multinomial and conditional logit, nested logit, multinomial probit, and random parameters logit. The last two models are estimated using simulation or Bayesian methods. For ordered data, standard multinomial models are ordered logit and probit, or count models are used if ordered discrete data are actually a count.

Keywords

Additive random utility model (ARUM); Binary outcomes; Categorical data; Categorical outcome models; Choice-based sampling; Cumulative distribution function (CDF); Discrete outcome models: *see* categorical outcome models; Heteroskedasticity; Limited dependent variable models; Logit models; Maximum likelihood; Maximum score methods; Multinomial models; Probit models; Qualitative response models: *see* categorical outcome models; Random parameters logit model; Semiparametric estimation; Simulation-based estimation; Tobit models

JEL Classifications
C25

Categorical outcome models are regression models for a dependent variable that is a discrete variable recording in which of two or more categories, usually mutually exclusive, an outcome of interest lies.

Categorical outcome models are also called discrete outcome models or qualitative response models, and are examples of a limited dependent variable model. Different models specify different functional forms for the probabilities of each category. These models are binomial or multinomial models, usually estimated by maximum likelihood.

Key early econometrics references include McFadden (1974), Amemiya (1981), Manski and McFadden (1981) and Maddala (1983). For textbook treatments see Amemiya (1985), Wooldridge (2002), Greene (2003) and Cameron and Trivedi (2005). The recent econometrics literature has focused on semiparametric estimation (see Pagan and Ullah 1999) and on simulation-based estimation of multinomial models (see Train 2003).

Binary Outcomes: Logit and Probit Models

Binary outcomes provide the simplest case of categorical data, with just two possible outcomes. An example is whether or not an individual is employed and whether or not a consumer makes a purchase.

For binary outcomes the dependent variable y takes one of two values, for simplicity coded as 0 or 1. If $y_i = 1$ with probability p_i, then necessarily $y_i = 0$ with probability $1 - p_i$, where i denotes the i^{th} of N observations. Regressors $\mathbf{x}_i$ are introduced by parameterizing the probability p_i, with

$$p_i = \Pr[y_i = 1 | \mathbf{x}_i] = F(\mathbf{x}_i'\beta),$$

where $F(\cdot)$ is a specified function and a single-index form is assumed.

The obvious choice of $F(\cdot)$ is a cumulative distribution function (CDF) since this ensures that $0 < p_i < 1$. The two standard models are the logit model with $p_i = \Lambda(\mathbf{x}_i'\beta) = e^{\mathbf{x}_i'\beta}/(1 + e^{\mathbf{x}_i'\beta})$, where $\Lambda(z) = e^z/(1 + e^z)$ is the logistic CDF, and the probit model with $p_i = \Phi(\mathbf{x}_i'\beta)$, where $\Phi(\cdot)$ is the standard normal CDF.

Interest usually lies in the marginal effect of a change in regressor on the probability that $y = 1$. For the r^{th} regressor, $\partial p_i/\partial x_{ir} = F'(\mathbf{x}_i'\beta)\beta_r$ where F' denotes the derivative of F. The sign of β_r gives the sign of the marginal effect, if F is a continuous CDF since then $F' > 0$, though the magnitude depends on the point of evaluation $\mathbf{x}_i$. Common methods are to report the average marginal effect over all observations or to report the marginal effect evaluated at $\bar{\mathbf{x}}$.

Parameter estimates are usually obtained by maximum likelihood (ML) estimation. Given p_i, the density can be conveniently expressed as $f(y_i) = p_i^{y_i}(1 - p_i)^{1-y_i}$. On the assumption of independence over i, the resulting log-likelihood function is

$$\ln L(\beta) = \sum_{i=1}^{N} \{y_i \ln F(\mathbf{x}_i'\beta) + (1 - y_i)\ln(1 - F(\mathbf{x}_i'\beta))\}.$$

It can be shown that consistency of the ML estimator requires only that $p_i = F(\mathbf{x}_i'\beta)$, that is, that the functional form for the conditional probability is correctly specified.

There is usually little difference between the predicted probabilities obtained by probit or logit, except for very low and high probability events. For the logit model $\ln[p_i/(1 - p_i)] = \mathbf{x}_i'\beta$, so that β_r gives the marginal effect of a change in x_{ir} on the log-odds ratio, a popular interpretation in the biostatistics literature.

A simpler method for binary data is OLS regression of y_i on $\mathbf{x}_i$, with White heteroskedastic robust standard errors used to control for the intrinsic heteroskedasticity in binary data. A serious defect is that OLS permits predicted probabilities to lie outside the (0,1) interval. But it can be useful for exploratory analysis, as OLS coefficients can be directly interpreted as marginal

effects and standard methods then exist for complications such as endogenous regressors.

When one of the outcomes is uncommon, surveys may over-sample that outcome. For example, a survey of transit use may be taken at bus stops to over-sample bus riders. This is a leading example of choice-based sampling. Standard ML estimators are inconsistent and instead one must use alternative estimators such as appropriately weighted ML.

The preceding discussion presumes knowledge of F. A considerable number of semiparametric estimators that provide consistent estimates of β given unknown F have been proposed. Manski's (1975) smooth maximum score estimator was a very early example of semiparametric estimation.

Index Models

Define a latent (or unobserved) variable y_i^* that measures the propensity for the event of interest to occur. If y_i^* crosses a threshold, normalized to be zero, then the event occurs and we observe $y_i = 1$ if $y_i^* > 0$ and $y_i = 0$ if $y_i^* \leq 0$. If $y_i^* = \mathbf{x}_i'\beta + u_i$, then

$$p_i = \Pr\left[y_i^* > 0\right] = \Pr\left[-u_i < \mathbf{x}_i'\beta\right] = F\left(\mathbf{x}_i'\beta\right),$$

where $F(\cdot)$ is the CDF of $-u_i$.

The logit model arises if u_i has the logistic distribution. The probit model arises if u_i has the more obvious standard normal distribution, where imposing a unit error variance ensures model identification. The probit model ties in nicely with the Tobit model, where more data are available and we actually observe $y_i = y_i^*$ when $y_i^* > 0$. And it extends naturally to ordered multinomial data.

Random Utility Models

In many economics applications the binary outcome is determined by individual choice, such as whether or not to work. Then the outcome should be the alternative with highest utility. The additive random utility model (ARUM) specifies the utility for individual i of alternative j to be $U_{ij} = \mathbf{x}_{ij}'\beta_j + \varepsilon_{ij}, j = 0, 1$, where the error term captures factors known by the decision-maker but not the econometrician. Then

$$\begin{aligned} p_i &= \Pr[U_{i1} > U_{i0}] \\ &= \Pr\left[(\varepsilon_{i0} - \varepsilon_{i1}) \leq \mathbf{x}_{i1}'\beta_1 - \mathbf{x}_{i0}'\beta_0\right] \\ &= F\left(\mathbf{x}_{i1}'\beta_1 - \mathbf{x}_{i0}'\beta_0\right) \end{aligned}$$

where F is the CDF of $(\varepsilon_{i0} - \varepsilon_{i1})$. For components x_{ir} of $\mathbf{x}_i$ that vary across alternatives (so $x_{i0r} \neq x_{i1r}$) it is common to restrict $\beta_{0r} = \beta_{1r} = \beta_r$. For components x_{ir} of $\mathbf{x}_i$ that are invariant across alternatives (so $x_{i0r} = x_{i1r}$) only the difference $\beta_{1r} - \beta_{0r}$ is identified.

The probit model arises, after rescaling, if ε_{i0} and ε_{i1} are i.i.d. standard normal. The logit model arises if ε_{i0} and ε_{i1} are i.i.d. type 1 extreme value distributed with density $f(\varepsilon) = e^{-\varepsilon}\exp(-e^{-\varepsilon})$. The latter less familiar distribution provides more tractable results when extended to multinomial models.

Multinomial Outcomes

Multinomial outcomes occur when there are more than two categorical outcomes. With m outcomes the dependent variable y takes one of m mutually exclusive values, for simplicity coded as $1, \ldots, m$. Let p_j denote the probability that the j^{th} outcome occurs. The multinomial density for y can be written as $f(y) = \prod_{j=1}^{m} p_j^{y_j}$ where $y_j, j = 1, \ldots, m$, are m indicator variables equal to 1 if $y = j$ and equal to 0 if $y \neq j$. Introducing a further subscript for the i^{th} individual and assuming independence over i yields log-likelihood

$$\ln L(\beta) = \sum_{i=1}^{N}\sum_{j=1}^{m} y_{ij} \ln p_{ij},$$

where the probabilities p_{ij} are modelled to depend on regressors and unknown parameters β.

There are many different multinomial models, corresponding to different parameterizations of p_{ij}.

Unordered Multinomial Models

Usually the outcomes are unordered, such as in choice of transit mode to work. The benchmark model for unordered outcomes is the multinomial logit model. When regressors vary across alternatives (such as prices), the conditional logit (CL) model specifies $p_{ij} = e^{\mathbf{x}_{ij}'\beta} / \sum_{k=1}^{m} e^{\mathbf{x}_{ij}'\beta}$. If regressors are invariant across alternatives (such as gender), the multinomial logit (MNL) model specifies $p_{ij} = e^{\mathbf{x}_i'/\beta_j} / \sum_{k=1}^{m} e^{\mathbf{x}_i'\beta_k}$, with a normalization such as $\beta_1 = 0$ to ensure identification. In practice some regressors may be a mix of invariant and varying across alternatives; such cases can be reexpressed as either a CL or MNL model.

The CL and MNL models reduce to a series of pairwise choices that do not depend on the other choices available. For example, the choice between use of car or red bus is not affected by whether another alternative is a blue bus (essentially the same as the red bus). This restriction, called the assumption of independence of irrelevant alternatives, has led to a number of alternative models.

These models are based on the ARUM. Suppose the j^{th} alternative has utility $U_{ij} = \mathbf{x}_{ij}'\beta + \varepsilon_{ij}, j = 1, \ldots, m$. Then

$$\begin{aligned} p_{ij} &= \Pr\left[U_{ij} \geq U_{ik} \text{ for all } k\right] \\ &= \Pr\left[\left(\varepsilon_{ik} - \varepsilon_{ij}\right) \leq \left(\mathbf{x}_{ij}'\beta - \mathbf{x}_{ik}'\beta\right) \forall\, k\right]. \end{aligned}$$

The CL and MNL models arise if the errors ε_{ij} are i.i.d. type 1 extreme value distributed. More general models permit correlation across alternatives j in the errors ε_{ij}.

The most tractable model with error correlation is a nested logit model. This arises if the errors are generalized extreme value distributed. This model is simple to estimate but suffers from the need to specify a particular nesting structure.

The richer multinomial probit model specifies the errors to be m–dimensional multivariate normal with $(m + 1)$ restrictions on the covariances to ensure identification. In practice it has proved difficult to jointly estimate both β and the covariance parameters in this model. A recent popular model is the random parameters logit model. This begins with a multinomial logit model but permits the parameters β to be normally distributed. For these two models there is no closed form expression for the probabilities and estimation is usually by simulation methods or Bayesian methods.

Ordered Multinomial Models

In some cases the outcomes can be ordered, such as health status being excellent, good, fair or poor.

The starting point is an index model, with single latent variable, $y_i^* = \mathbf{x}_i'\beta + u_i$. As y^* crosses a series of increasing unknown thresholds we move up the ordering of alternatives. For example, for $y^* > \alpha_1$ health status improves from poor to fair, for $y^* > \alpha_2$ it improves further to good, and so on. For the ordered logit (probit) model the error u is logistic (standard normal) distributed.

An alternative model is a sequential model. For example, one may first decide whether or not to go to college ($y = 1$) and if chose college then choose either two-year college ($y = 2$) or four-year college ($y = 3$). The two decisions may be modelled as separate logit or probit models.

A special case of ordered categorical data is a count, such as number of visits to a doctor taking values 0, 1, 2, …. An ordered model can be applied to these data, but it is better to use count models. The simplest count model is Poisson regression with exponential conditional mean $\mathrm{E}[y_i | \mathbf{x}_i] = \exp\left(\mathbf{x}_i'\beta\right)$. Common procedures are to use the Poisson but obtain standard errors that relax the Poisson restriction of variance-mean equality, to estimate the richer negative binomial model, or to estimate hurdle or two-part models or with-zeroes models that permit the process determining zero counts to differ from that for positive counts.

Multivariate Outcomes and Panel Data

Multivariate discrete data arise when more than one discrete outcome is modelled. The simplest example is bivariate binary outcome data. For example, we may seek to explain both employment status (work or not work) and family status (children or no children). The standard model is a bivariate probit model that specifies an index model for each dependent variable with normal errors that are correlated. Such models can be extended to permit simultaneity.

For panel binary data the standard model is an individual specific effects model with $p_{it} = F\left(\alpha_i + \mathbf{x}_{it}'\beta\right)$ where α_i is an individual specific effect. The random effects model usually specifies $\alpha_i \sim N\left[0, \sigma_\alpha^2\right]$ and is estimated by numerically integrating out α_i using Gaussian quadrature. The fixed effects model treats α_i as a fixed parameter. In short panels with few time periods consistent estimation of β is possible in the fixed effects logit but not the fixed effects probit model. If $\mathbf{x}_{it}$ includes $y_{i,\,t-1}$, a dynamic model, fixed effects logit is again possible but requires four periods of data.

See Also

- ▶ Contingent Valuation
- ▶ Hierarchical Bayes Models
- ▶ Logit Models of Individual Choice
- ▶ Maximum Score Methods
- ▶ Semiparametric Estimation
- ▶ Simulation-Based Estimation

Bibliography

Amemiya, T. 1981. Qualitative response models: A survey. *Journal of Economic Literature* 19: 1483–1536.

Amemiya, T. 1985. *Advanced econometrics*. Cambridge, MA: Harvard University Press.

Cameron, A., and P. Trivedi. 2005. *Microeconometrics: Methods and applications*. Cambridge: Cambridge University Press.

Greene, W. 2003. *Econometric analysis*. 5th ed. Upper Saddle River: Prentice-Hall.

Maddala, G. 1983. *Limited-dependent and qualitative variables in econometrics*. Cambridge: Cambridge University Press.

Manski, C. 1975. The maximum score estimator of the stochastic utility model of choice. *Journal of Econometrics* 3: 205–228.

Manski, C., and D. McFadden, ed. 1981. *Structural analysis of discrete data with econometric applications*. Cambridge, MA: MIT Press.

McFadden, D. 1974. Conditional logit analysis of qualitative choice behavior. In *Frontiers in econometrics*, ed. P. Zarembka. New York: Academic Press.

Pagan, A., and A. Ullah. 1999. *Nonparametric econometrics*. Cambridge: Cambridge University Press.

Train, K. 2003. *Discrete choice methods with simulation*. Cambridge: Cambridge University Press.

Wooldridge, J. 2002. *Econometric analysis of cross section and panel data*. Cambridge, MA: MIT Press.

C

Catholic Economic Thought

Pedro Teixeira and António Almodovar

Abstract

Although Catholic economics' roots date back to the beginnings of Christianity, its emergence as a structured discourse developed later and slowly. The establishment of a distinctive Catholic approach to modern social and economic problems had to await a more extensive development of the market system and the emergence of political economy. The most prolific period for Catholic economic thought began in 1891 and continued until the end of the Second World War. In the second half of the twentieth century the church's interest focused on the analysis of such themes as development, international aid and cooperation.

Keywords

Capitalism; Catholic economic thought; Charitable contributions; Class; Concentration of wealth; Economic freedom; Exploitation; Fair trade; Foreign aid; Just wage; Natural law; Poverty; Redistribution of wealth; *Rerum Novarum*; Salamanca School; Social justice;

Social responsibility; Socialism; Solidarity; Subsidiarity; Usury

JEL Classifications
B5

Catholic economic thought is the outcome of a series of efforts to evaluate the workings of economic life according to a definite set of religious principles. In its more evolved forms, these efforts have inevitably led to include the findings of political economy, and later of economics, in its assessment of economic life, but also to assess the findings of economic analysis itself. According to a strict ecclesiological perspective, only the hierarchy of the Church is authorized to identify the appropriate religious principles that are to be applied to the analysis of the livelihood of man. Therefore, some of the assessments made by Catholics may be considered by the Church's hierarchy as inappropriate.

Catholic economic thought is not to be confused with the social doctrine of the Catholic Church. Since 1891, the most relevant religious principles for the appraisal of social questions from a theological perspective are gathered in the social doctrine of the Church, which is essentially based in the so-called social encyclicals, which are official documents written by several popes, often based on documents prepared by other high-ranking Church officials. These documents emerged as attempts to offer a better moral and philosophical framework for the workings of a modern society, not as in-depth and systematic discussions of man's economic life or as blueprints for a thorough discussion of economic concepts and theories. By being focused on the material aspects of life, Catholic economic thought is prone to give more emphasis to particular problems – such as usury and finance, social and labour questions or, later, the outline of an alternative economic and social system. However, Catholic economic literature has as a rule been less focused than political economy on technical aspects.

Catholic economic thought has an inescapable doctrinal and normative accent. Its '*ought*' sentences are considered as quasi-*positive* ones, in the sense that they were allegedly meant by God to become factual statements in a society functioning in accordance with natural law (see Barrera 2001, pp. 117–31). This normative stance acts as an explicit incentive for social action, in order both to amend the workings of existing institutions and to establish new ones – such as charitable institutions, cooperatives, institutions of mutual assistance, and particular ways of labour–capital association. In certain periods, when Catholics were more openly engaged in the revision of economic life, their thought went as far as to suggest the establishment of a specific economic system, which was a third way between the liberal and the socialist ones. But, even when they were more focused on the implementation of particular social and economic measures, people engaged in these initiatives also left some thoughts that are of more general interest.

Early Attempts to Formulate Catholic Economic Thought

Although the roots of Catholic economics date back to the beginnings of Christianity, the emergence of a structured discourse developed later and slowly. Thus, even if some of the basic Catholic principles for social and ethical teaching were already present in the gospels and in the patristic literature, the systematic theology of Aquinas was instrumental in the move towards a more organized approach to economic problems. The earliest scholarly attempt to produce an explicit and meaningful set of theological principles applied to economic problems was performed in the sixteenth century by authors belonging to the Salamanca School. Under the philosophical umbrella provided by Thomism, Dominicans like Vitoria, Soto, and Mercado, and Jesuits like Molina, Mariana and Lugo addressed the problems of usury, prices, and justice in wages. Although these ideas were not formally adopted by the Church, this literature was widely used by confessors in search of appropriate answers for the moral questions raised by the development of economic activity (on the economic thought of the school of Salamanca, see Grice-Hutchinson 1978, 1993,

and Camacho 1998; these authors are also relevant as examples of a revival of Thomist moral theology, which they applied to international law: see Curran 2002).

The Nineteenth Century

The establishment of a distinctive and clear-cut Catholic approach to modern social and economic problems had nevertheless to wait for a more extensive development of the market system and the emergence of political economy. By the late 1830s, the first Catholic political economists were already trying to infuse some basic Christian values into the teachings of classical political economy. Together with the socialists, they were concerned about the consequences of unbridled competition, the concentration of riches in the hands of the few, the exploitation of the poor and weak, and the existence of pervasive unemployment. However, contrary to socialists, Catholics thought that those evils, together with excessive materialism and burgeoning social and political unrest, were to be curbed by individuals renouncing material goods and by extended charity, not by abolishing private property or an expansion of the state. Their criticism voiced the fundamental Christian values of universal fraternity and respect for human dignity, as expressed in the Gospels and in the Apostolic letters.

It is important to note that in the mid-nineteenth century there was a series of authors who wrote on economic subjects from a Catholic perspective before the *Rerum Novarum*, the encyclical of Pope Leo XIII on capital and labour, promulgated 15 May 1891. Among these we find the names of Charles de Coux, Alban de Villeneuve-Bargemont, Joseph Droz, Charles Périn, and Matteo Liberatore. The first four authors are representative of the Catholic perspectives that emerged gradually in the context of nineteenth-century France and Belgium. Three of them – Coux, Périn and Droz – were openly against any solution for economic problems that would require increased state intervention, and they asked the rich to voluntarily avoid all extreme forms of exploitation and competition; as a rule, they were reasonably sympathetic towards political economy, and may be considered as the forerunners of the conservative tendency that was later organized around the Angers school. Villeneuve-Bargemont had less confidence in voluntary individual action as a remedy for the emerging poor question. Contrary to the Catholic *conservative* approach, Villeneuve-Bargemont thought that the scale of the problem was so serious that the state should intervene in favour of the labouring masses before they fell irrevocably under the spell of socialism. Thus he may be considered as a precursor of the so-called *progressive* tendency, later developed by the Fribourg Union and the Liège school. Matteo Liberatore deserves mention, since he was one of the persons involved in the drafting of Leo XIII's *Rerum Novarum* (1891). His views were closer to Villeneuve-Bargemont than to Charles Périn, since he believed that modern poverty was a phenomenon that could not be solved by traditional means (charity), because its causes were embedded in modern social and economic organization. Modern exploitation and modern social unrest were seen not only as consequences of the acceptance of a social and economic model based on the erroneous philosophical notions underlying political economy, but also because the spread of the latter stimulated people to act in a way that damaged social cohesion. Contrary to materialist and utilitarian views, wealth should considered as a means, not an end, and should be distributed according to justice; and the human person should always be respected – meaning that in no circumstance should labour be considered as a mere commodity to be bought and sold in the market. Once individualism and competition were once again checked by an attention to mutual needs, modern phenomena such as the class struggle (between labour and capital) would vanish and a sense of mutually beneficial collaboration would take its place. Measures such as the re-establishment of updated medieval corporations – which had to be adapted to the new realities and not just re-established, were a possible institutional solution for bringing peace and harmony to the relations between producers, namely because they would help to resume natural

social relations and reduce the moral, social, and professional void in which liberalism had placed the individuals. Efforts to promote the modern resurgence of these institutions were at the origins of what later became corporatism (see below).

The Golden Age, 1891–1940s

The most prolific period for Catholic economic thought began in 1891 and continued up to the end of the Second World War. Stirred by Leo XIII's *Rerum Novarum*, the willingness to address social and economic questions gave rise to extensive debate and to intense publishing activity (see De Rosa 2004; Hobgood 1991, p. 112).

The central issue of *Rerum Novarum* is the condition of workers, especially industrial workers, and the moral and material risks arising from what was seen as their degrading situation. Leo XIII made clear from the outset that he considered the major cause to be the political and economic transformations of the previous hundred years. This had destroyed or seriously damaged valuable traditional social structures such as medieval corporations. It had also launched a process of secularization of the legal and political framework, which had greatly diminished the moral influence of the Church. Liberalism had created a social vacuum in which unregulated competition, greed and usury had prospered, resulting in a substantial concentration of wealth and power. The latter eventually created an unbalanced distribution of privileges that made possible the exploitation of the workers by the all-mighty owners of capital. Leo XIII also asserted that the supposed remedies offered by socialists were inadequate to the task. In addition to the obvious problem of atheism, the crucial issue in the Church's critique of socialism was the former's concern with private property. Although the Church criticized the extreme capitalist/individualist/ utilitarian uses of private property, these criticisms did not question its fundamental existence.

The Church proposed a new relationship between workers and capitalists. Workers should opt for non-violent ways of solving labour disputes, and should perform faithfully and completely the tasks that were allocated to them. In return, paramount among the duties of capitalists was the acknowledgment of and the respect for the human dignity of the workers. This meant respecting the workers' physical and intellectual health, and the payment of a fair family wage that would put a stop to the need for female and youth labour. Capitalists' social responsibility was central to the way Christians should relate to wealth. Leo XIII underlined the ephemeral and secondary nature of earthly wealth and success. If the Church accepted the inequality of property, it also cared for the poorest members of society, knowing that, unless these were actively supported, they would fall into a state of quasi-serfdom, which would lead to social disruption.

One of the outcomes of the *Rerum Novarum* was the development of an array of books, typically bearing the title of *Principles or Courses on Social Economics*. Often written by Jesuits for the use of both clergy and active Catholic laity, this peculiar type of book tried to re-embed the political economy into a social philosophy so as to secure a coherent and global society based on Christian values (Galindo 1996, p. 143). The authors of such works had to perform complex scholarly work if they were to fulfil their aim. First, they had to explain classical political economy to their readers; then they had to introduce and explain the Pope's criticisms of the philosophical tenets underlying economic liberalism; next they had to deal with socialism, in order to make sure that this doctrine would not be seen as a possible alternative to the shortcomings of economic liberalism; and finally, they had to highlight the proper course of Catholic thought and action that was to be followed in order to put right contemporary evils. Authors that engaged in this type of work include Charles Antoine, S.J. (1896), Giuseppe Toniolo (1907–9) and Heinrich Pesch, S.J. (1905–26), the latter being considered by Schumpeter as the best example of neo-scholasticism (1954, p. 765). Another set of books focused on the outlines of a specifically Catholic system – a third, neo-corporative, way between liberalism and socialism. This system had its roots both in France (Mun, La Tour du

Pin) and Germany (Vogelsang, Kettler), and was further developed under the auspices of the Liège School. It was eventually accepted, if not warmly supported, by the encyclical *Quadragesimo Anno*.

Quadragesimo Anno appeared in 1931, when Pius XI took the initiative in clarifying and updating the position of the Catholic Church on the economic and social condition of the contemporary world. His view was that, although capitalism per se was not an evil system, there was a problem with the way it had developed, for it had led to economic despotism, namely, a concentration of wealth which gave to a few members of society huge power, which was often used to influence and subjugate governments and countries. The subjugation of the state to the interests of a wealthy minority, whose power was nurtured by ambition, greed and speculative behaviour, fostered social disorder and could lead to the collapse of essential social bonds.

The Church supported the existence of private property, but it also underlined its dual nature (individual and social) and the difference between property ownership and property usage. Hence, the relations between capitalists and workers in the capitalistic system should be reorganized according to this view. According to Pius XI, labour and capital did have common interests, and this communality of efforts and purposes called for a sharing of both the responsibility for the productive process and of the wealth created, including the profits resulting from the productive activity. Commutative justice would be insufficient, and should be complemented by social justice.

Pius XI also emphasized the principle of subsidiarity. According to this principle, the state should not intervene when intermediate levels of society (associations, local community, and family) could act effectively. Social harmony ought therefore to be built upon the contribution of intermediate communities and groups, these taking multiple forms. However, the reconstruction of the social fabric, which had been ruined by unlimited competition and the concentration of wealth, required the state to regulate competition, subordinating it to the higher values of justice and charity. To accomplish the necessary rebalance of social power in order to promote the common good, Pius XI made explicit references to the advantages and risks of the emerging corporatist organization (in Italy and elsewhere). Overall, he thought that the advantages (pacifying society, curbing the insidious influence of socialist organizations, and bringing together workers and capitalists in the search for the common good) could outweigh the possible risks of bureaucratization and state dirigisme. Pius XI saw the establishment of the corporative system as a step in the right direction, towards a Christian social-economic order, through its contribution to a harmonious society and its emphasis on the pursuit of the common good.

The Post-war Period

At the beginning of the 1960s, the Catholic Church underwent profound institutional and theological changes. With the Second Vatican Council (1962–5), Thomism, the theological and philosophical basis of the earlier social and economic doctrine of the Church, lost its unique status (see Nichols 2002, pp. 139–43). Vatican II also marked a change in the role of the laity, and opened the dialogue between different churches. Although until the early 1960s, socialism and communism stood at the forefront of Church's criticisms, some bridges were later to be established with Marxist sociology (see Curran 2002, pp. 201, 203).

The Catholic Church's approach to economic problems also took a different direction in the second half of the twentieth century, now focusing in the analysis of themes like development and North–South relationships, international aid and cooperation. This is particularly visible in John XXIII's *Mater et Magistra* (1961) and Paul VI's *Populorum Progressio* (1967). In the latter, Paul VI considered that the wealthiest nations had the duties of solidarity, justice and charity towards less developed ones, and that these duties should be addressed through international aid, fair trade and a framework conducive to mutual progress. He was particularly critical of free trade, since he regarded any exchange between unequals as potentially unjust. Hence, he called for fair and just competition between nations.

The social question was nevertheless not forgotten. In the *Mater et Magistra*, John XXIII stated that wages should not be left to market forces alone, for they should be determined by the laws of justice and equity. Private property was not to be considered solely as a right that should be protected, but also as an obligation to practise solidarity among human beings. John XXIII also gave explicit support to the political organization of workers in order to promote their legitimate rights. This text was also the first to address, and largely support, the so-called welfare state and its associated system of social insurance and social security, on the grounds of its contribution to the desirable redistribution of wealth. Although the Church kept its distance vis-à-vis socialism, Paul VI considered that there were some possibilities for cooperation between Catholics and socialist movements insofar as this contributed to a more just society (see his apostolic letter *Octogesima Adveniens* on the occasion of the 80th anniversary of *Rerum Novarum*, 1971).

The dialogue between economic analysis and theology was, if not on hold, at least withdrawn to the backstage. This is likely to have been for several different reasons, ranging from the changing priorities in theology and a new emergence of ecclesiological concerns with the inner life of the Church, to the growing professionalization of economics, which made it ever more difficult to acquire the desirable proficiency in both fields (see Wilson 1997, pp. 88–9 and 113). In the late 1950s, Catholic writers like Achille Dauphin-Meunier and Jean-Yves Calvez had already begun to assert that the Church had no other wish than to present its own social doctrine. To these authors, the Church was not to offer or to support 'an economic theory' but only a 'philosophical and religious clarification of the fundamental aspects of human existence within economic relationships' (Calvez and Perrin 1958, p. 11).

The Contemporary Situation

Catholic social doctrine received a significant stimulus in the 1980s and 1990s with John Paul II. He used the 90th and 100th anniversaries of *Rerum Novarum* to express views on the economic realm. In *Laborem exercens* (1981) he focused on the role of work as a central feature of all human activity and therefore of all economic activity. He considered that contemporary developments in technological, economic and political conditions had reinforced the pastoral care that the Church should associate with all issues related to work, such as unemployment and lifelong learning. He criticized what he considered the error of considering human labour solely according to its economic purpose, and underlined the principle of priority of human labour over capital, which should not be attained through class or social warfare but by peaceful struggle for social justice. Likewise, in *Centesimus Annus* (1991) he focused on the harshness of the modern conditions of the working class and pointed out how erroneous the collectivist and totalitarian solution was. Thus he insisted on the idea of redistribution of wealth in order to fulfil 'the universal destination of material goods'. John Paul II also devoted special attention to economic and social development, with particular attention being paid to issues such as international division of labour, international debt and poverty. In the encyclical *Sollicitudo Rei Socialis* (1987) he criticized both 'liberal capitalism' and 'Marxism collectivism' and proposed a view of 'authentic human development' which was not only economic but also social and spiritual. Thus, underdevelopment had not only social and economic causes, but also moral ones, not the least being the lack of international solidarity that denied human interdependence beyond national or political borders. His position vis–àvis social warfare and any possible analytical or political convergence with Marxism is vividly illustrated by the reaction of the Church's hierarchy to Liberation Theology, whose main proponents were either silenced or led to abandon the Catholic Church because of the restrictions imposed on them regarding teaching, preaching and writing.

Modern Catholic theology has focused on achieving a comprehensive and coherent presentation of social ethics (see Curran 2002). Those who give a certain emphasis to economic aspects (see Barrera 2001; Hobgood 1991), always take

care to reiterate 'the caveat that [the Church's social teachings do] not offer an alternative school of thought between classical laissez-faire capitalism and socialist centralized planning' (Barrera 2001, p. viii). Notwithstanding this change of focus, the modern effort to systematize the teachings of the encyclicals has led in some cases to the identification of six basic principles: universal access, the primacy of labour, subsidiarity, socialization, solidarity, and stewardship (2001, p. 1, and table on p. 258). By means of these principles, the criticisms addressed to economics continue to stress its defective philosophical base and go on emphasizing the collective risks that are incurred by a society unwilling to restrain excessively individualist, materialist, and utilitarian behaviour. The claims of contemporary Catholic economic thought therefore continue to emphasize the need for justice and equity, something that can be achieved only through the establishment of corrective measures to the workings of the market in order to prevent its deleterious action on the social fabric. The basic appeal therefore remains, that economics should not refuse the normative approach provided by the Catholic view of mankind.

See Also

- ▶ Aquinas, St Thomas (1225–1274)
- ▶ Ethics and Economics
- ▶ Religion and Economic Development
- ▶ Scholastic Economics

Bibliography

Barrera, A.O.P. 2001. *Modern Catholic social documents and political economy*. Washington, DC: Georgetown University Press.

Calvez, J.-Y., and J. Perrin. 1958. *Église et Société Économique. L'énseignement social des Papes de Léon XIII à Pie XII*. Paris: Éditions Montaigne.

Camacho, F.G. 1998. *Economia y Filosofia Moral. La formación del pensamiento económico europeo en la Escolástica española*. Madrid: Editorial Sintesis.

Curran, C.E. 2002. *Catholic social teaching, 1891–present: A historical, theological and ethical analysis*. Washington, DC: Georgetown University Press.

De Rosa, G. 2004. *I tempi della 'Rerum Novarum'*. Rome: Rubbettino/Istituto Luigi Sturzo.

Galindo, A. 1996. *Moral Socioeconómica*. Madrid: Biblioteca de Autores Cristianos.

Grice-Hutchinson, M. 1978. *Early economic thought in Spain, 1177–1740*. London: George Allen & Unwin.

Grice-Hutchinson, M. 1993. *Ensayos sobre el pensamiento económico en España*. Madrid: Alianza Universidad.

Hobgood, M.E. 1991. *Catholic social teaching and economic theory: Paradigms in conflict*. Philadelphia: Temple University Press.

Nichols, A.O.P. 2002. *Discovering Aquinas: An introduction to his life, work and influence*. London: Darton, Longman & Todd.

Schumpeter, J.A. 1954. *A history of economic analysis*, 1994. London: Routledge.

Wilson, R. 1997. *Economics, ethics and religion: Jewish, Christian and Muslim economic thought*. London: Macmillan.

Cattaneo, Carlo (1801–1861)

R. P. Bellamy

Cattaneo was a leading spokesman for social and political reform in his native Lombardy. A polymath, he made important contributions to history, geography, linguistics and philosophy and took a prominent role in politics, as well as writing on economics and engaging in various business ventures. However, he preferred the title of economist to all others, and a concern with economic reform runs through his work. An admirer of Charles Bonet, a follower of Condillac, he developed his own theory of human progress from barbarism to civility. At the heart of his thesis was a sensationalist epistemology adapted from Vico, which he called the psychology of associated minds. He argued that if individuals were allowed to experience sufficient contrasting ideas and situations then humankind would gradually improve and both our needs and the means of satisfying them infinitely multiply. He was therefore a staunch advocate of both political liberty and free trade, which he regarded as linked. He criticized the feudal privileges and economic nationalism of the period, calling for the abolition

of the decrees against the Jews and opposing the protectionist doctrines of Friedrich List, but defended private property as an inalienable right essential to individual liberty and vehemently attacked socialist proposals for public ownership, especially Proudhon's.

He regarded the contrast between the highly developed Lombard agriculture, based on irrigation schemes devised by small proprietors, and the backward cultivation of the vast feudal estates in the south as illustrating the links between liberalism and economic development. He believed the collective enterprises of Lombard farmers, based on mutual self-interest, provided a model for republican self-government, but were only feasible within small territories where a relative homogeneity of interest and culture obtained. He therefore attacked the projects of both Cavour and Mazzini for Italian unification under a single government, devising an alternative proposal for a federal republic. He felt federalism provided the best antidote to the centralizing tendencies of the age, and ultimately hoped for a United States of Europe on the North American model.

His ideas, largely disseminated through reviews such as *Politecnico* (which he founded in 1839 and ran from 1839 to 1844 and from 1860 to 1863), were very influential at the time, and he was even asked by Gladstone to devise a scheme for agricultural reform in Ireland. He took a major part in the development of the Italian railway network, arguing that the lines should be constructed according to economic rather than political considerations. Participation in the revolution of 1848 against Austrian rule led to his exile in Switzerland. Although Garibaldi and Cavour separately sought his advice and aid in 1860, his federal republicanism led him to fall out with both of them and he was excluded from academic and political posts in the new Italian kingdom.

Selected Works

1958. In *Scritti Economici*, 3 vols, ed. A. Levi. Florence: Le Monnier.
1960. In *Scritti Storici e Geografici*, 4 vols, ed. G. Salvemini and E. Sestan. Florence: Le Monnier.
1964. In *Scritti Politici*, 4 vols, ed. M. Boneschi. Florence: Le Monnier.

References

Ambrosoli, L. 1960. *La Formazione di Carlo Cattaneo*. Milan/Naples: Ricciardi.
Greenfield, K.R. 1965. *Economics and Liberalism in the Risorgimento: A study of nationalism in Lombardy 1814–48*, Rev. ed. Baltimore: Johns Hopkins Press.
Lovett, C.M. 1972. *Carlo Cattaneo and the politics of the Risorgimento*. The Hague: Nijhoff.

Causal Inference

C. W. J. Granger

When a particular event is observed, such as an economic variable taking a value in some region of the set of all possible values, it is natural to ask why that event occurred rather than some other. If, just earlier, some other event was observed to occur, it is also natural to ask if the joint observation of the two events indicates a relationship and possibly one that could be called an influence of one event by another, or even a causation. For a unique, or very rare event, such as the start of a world war, it will be very difficult to present more than sensible and suggestive statistical evidence about causation. However, in economics, values for many variables are observed with great regularity, such as daily stock market prices or monthly production figures and so a generating mechanism can be postulated that produces these values and the investigation and understanding of this mechanism is obviously one of the main tasks for the economist. In such studies, ideas such as theories, laws and causation arise very naturally, and economists in their workings use such words very frequently. It is unfortunately true that not all writers give the same meanings to these words. The understanding of causality is not the same for all economists, but this is hardly surprising as statisticians and philosophers are also not in agreement among themselves.

Economists who have attempted to discuss the meaning of causation in economics include Herbert Simon (1953), Herman Wold (1954), Julian Simon (1970), Sir John Hicks (1979), and Arnold Zellner (1979). Most of the writers emphasize the difference between a mere association and the deeper sub-class of associations that might be called causal relationships. To distinguish these, some statisticians have emphasized the use of experimental studies, but these are rarely available in economics and so this aspect of causation will not be further considered.

One can either discuss causation in very general, abstract terms or the discussion can be focused on the specific question of whether it is possible to test for causation using the data available. The latter requires an operational procedure and definition. There are basically two types of causal testing situations. In the first, a population of economic agents is observed and some variables measured for each, for example, the amount of electricity used by a household. The totality of these measurements gives a distribution. A question can then be asked – why does this household use more electricity than that one? This is a cross-sectional causality question. It is also possible to measure parameters of the distribution, such as the mean or the variance, and to ask why these parameters are changing through time. Thus, the question is asked, why is electricity demand higher this year than last? This could be called a temporal causality question. The definitions of causality and their interpretations may differ between these two cases.

It is convenient to assume the existence of a quantity called the 'degree of belief' held by an individual about the correctness of some causal theory or proposition and to assume further that this quantity can be represented as a probability. The objective of any causal analysis, such as a statistical test, might be to try to influence the degree of belief of oneself or of others. For this purpose, the analysis need not be complete or perfect, but merely to have enough value to make one reconsider one's beliefs.

A mere association between a pair of economic variables, such as a correlation or a non-independent joint distribution, is insufficient to determine a causation, partly because such associations are symmetric between the variables, the extent to which X is correlated to Y, or can be explained by Y, is exactly the same as Y is correlated, or explained, by X. It is generally thought that causation is a non-symmetric relationship, and there are various ways in which asymmetry can be introduced, the most important of which are controllability, a relevant theory, outside knowledge, and temporal priority. Amongst the economic writers, each has its advocates and detractors.

Concerning controllability, Strotz and Wold (1960) write:

> z is a cause of y if, by hypothesis, it is or 'would be possible' by controlling z indirectly to control y, at least stochastically. But it may not be possible by controlling y indirectly to control z this way.

Essentially this idea is from their experimental background and uses hypothetical experiments. By utilizing enough knowledge about lack of controllability in a system, so that some possible causal links are put to zero, tests can be constructed on the remaining links. This would obviously be the case if a system of variables, all measured at the same time, could be displayed recursively, so that the jth equation involved only the first j variables. However, by redefining variables as linear combinations of the original set, such a recursive system can always be achieved and not uniquely, unless there are sufficient identifying qualifications on the system. J. Simon suggests that controllability is required to make causal analysis useful for policy-makers. The equivalence of causation and controllability is not generally accepted, the latter being perhaps a deeper relationship. If a causal link were found and was not previously used for control, the action of attempting to control with it may destroy the causal link.

Hicks, Zellner and J. Simon, in discussing causal links, all emphasize the relevance of a sound economic theory. Hicks (1979) accepts static or equilibrium theory as sufficient for use, while J. Simon (1970) suggests that a statement that is 'logically connected to the general framework of systematic economics is much more

likely to be considered causal than one that stands alone'. Thus, the theory is used to increase the degree of belief, and these writers suggest that a strong degree of belief cannot be achieved without a convincing theory. Zellner (1979) takes a much stronger view, leaning heavily on the work of the philosopher H. Feigl who says that 'the clarified (or purified) concept of causation is defined in terms of *predictability according to a law* (or more adequately, according to a set of laws)'. In his work, Zellner appears to be saying that for him a degree of belief cannot be anything but very small unless the causal analysis is based on some generally acceptable economic theory. He gives no examples of such economic laws and it is interesting to note that Hicks (1979, p. 2) says that 'there are few economic laws that are at all firmly based'.

Concerning temporal priority, it is generally, although not universally, accepted that the cause cannot occur after the effect. It is also frequently assumed that the cause will occur before the effect, providing a convenient asymmetry, but this view is certainly more controversial. Both Zellner and Hicks firmly reject it and Hicks maintains that instantaneous and contemporaneous causality is the 'characteristic form of the causal relation in modern economics'. It is certainly true that much economic theory is written as though causation is instantaneous. However, as Hicks also points out, all economic variables are accumulations of the outcomes of economic decisions and it is difficult to present a sensible decision mechanism in which there is an instantaneous relationship between the observed inputs to the decision (the causes) and the observed outputs (the effects). Thus, for statistical testing purposes, which has to use just observed variables, the temporal priority assumption appears to be more reasonable. It is also clear that if any part of the cause cannot occur later than any part of the variable being effected, instantaneous causation cannot occur between some pairs of stock and flow variables. For example, production of steel in a month could not instantaneously cause production of automobiles in the month, as part of one variable occurs after part of the other. There is always the possibility of apparent instantaneous causation occurring because of temporal aggregation or missing common causes.

Occasionally other outside information is used to break the symmetry of association. One variable may be thought to be generated outside the economic system, such as a weather variable, so that causation can only flow from it to part of the economy. This idea is the classical one of exogeneity. For a discussion of this topic with generalizations concerned with estimation problems, see Engle et al. (1983). A particular case is when a variable is thought to be completely controlled, such as tax rates or possibly money supply, so that controlled money could cause price changes but not vice versa. In all these cases, the outside information may be useful, if it is correct.

Although many important economic questions can be phrased in the cross-sectional causal situation, they have received little causal testing in that context, except under the 'outside information' assumption. However, many tests have been conducted for economic questions that can be stated as temporal causation. These tests have been conducted using the concepts known in the literature as 'Granger-causation'. This approach is based on two axioms – that the cause will occur before the effect (strict temporal priority) and that the cause contains unique information about the effect. The second can be stated more formally as follows. Let A_t represent all the observable information available at time t and $A_t - Y_t$ represent all this information except that contained in the series $Y_{t-j}, j \geq 0$. Then Y_t will be said to cause X_{t+1} if

$$\Pr \text{ob}(X_{t+1} \text{in} C | A_t) \neq \Pr \text{ ob}(X_{t+1} \text{in} C | A_t - Y_t)$$

for any region C. The two axioms have the simple consequence that any well-behaved function $f(X_{t+1})$ will be generally better forecast using any cost function as a criterion. Thus, tests of this type of causation potentially can be based on forecastability but to be operational some simplifications are required. If one has a belief about a temporal causation then it could be called a prima facie causality. If a test is based on the above definition, but with the unuseful universal information-set replaced by a restricted but practical information set I_t, and if the test finds evidence

for causation, then the relationship remains a prima facie cause. The set I_t will consist of a group of time series and the larger and more relevant it is, the more stringent will be the test; it is then more likely that degrees of belief will change. The choice of the causation to investigate and the choice of I_t will probably depend on some theory, but this could be a low-level theory and, if the tests so suggest, may be worth further development. In practice, tests are rarely based on distributions but on parameters of the distributions such as means. This could be stated as 'Y_t is a prima facie cause in mean of X_{t+1} with respect to I_t' if

$$E[X_{t+1}|I_t] \neq E[X_{t+1}|I_t - Y_t].$$

It will follow that X_{t+1} is better forecast, using a least squares criterion, if Y_t is used than if it is not used. Standard time-series modelling techniques will provide models of X_{t+1} based on I_t and on $I_t - Y_t$ and the post-sample forecasting ability of the two models can then be used to test this particular form of causation. Some of these tests are described in Pierce and Haugh (1977) and evaluated in Nelson and Schwert (1982). They are generally linear in data, although do not have to be, and, if misapplied, can of course lead to incorrect results. To correspond strictly to the definition, tests should be based on the post-sample forecasting abilities of the alternative models.

The definition has both some advantages and some problems, and these are discussed in Granger (1980). In theory, the tests are not altered if backward filters are applied to the data, but some kinds of seasonal adjustments or measurement errors can give problems. If Y_t causes X_{t+1} the X_t may, but need not cause Y_{t+1}, so that feedback can occur but need not. Similarly, if Y_t causes X_{t+1} and X_t causes Z_{t+1} then Y_t may, but need not cause, Z_{t+2}. It has to be remembered when interpreting results based on tests that missing common causal variables can always alter the interpretation, that causation may be lost if one variable is controlled so as to reduce the strength of the causal link, and that temporal aggregation or using data measured over intervals much wider than actual causal lags can also destroy causal interpretation.

See Also

- ▶ Autoregressive and Moving-Average Time-Series Processes
- ▶ Causality in Economic Nodels
- ▶ Endogeneity and Exogeneity
- ▶ Spectral Analysis
- ▶ Stationary Time Series
- ▶ Time Series Analysis

Bibliography

Engle, R.F., D.F. Hendry, and J.F. Richard. 1983. Exogeneity. *Econometrica* 51: 277–304.

Granger, C.W.J. 1980. Tests for causation – A personal viewpoint. *Journal of Economic Dynamics and Control* 2: 329–352.

Hicks, Sir J. 1979. *Causality in economics*. New York: Basic Books.

Nelson, C.R., and G.W. Schwert. 1982. Tests for predictive relationships between time series variables. *Journal of the American Statistical Association* 77: 11–18.

Pierce, D.A., and L.D. Haugh. 1977. Causality in temporal systems. *Journal of Econometrics* 5: 265–293.

Simon, H. 1953. Causal ordering and identifiability. In *Studies in econometric method*, Cowles Commission Monograph No. 14, ed. W.C. Hood and T.C. Koopmans. New York: John Wiley.

Simon, J.L. 1970. The concept of causality in economics. *Kyklos* 23: 226–252.

Strotz, R.H., and H. Wold. 1960. Recursive versus nonrecursive systems: An attempt at synthesis. *Econometrica* 28: 417–427.

Wold, H. 1954. Causality and econometrics. *Econometrica* 22: 162–177.

Zellner, A. 1979. Causality and econometrics, policy and policy making. *Carnegie-Rochester Conference Series on Public Policy* 10: 9–54.

Causality in Economic Nodels

Herbert A. Simon

Causal notions arise when we seek to understand the workings of a complex system by analysing it into component subsystems and mechanisms. Thus, if we wish to understand the quantities of strawberries that are produced and consumed and

the prices at which they are exchanged, we may consider a number of mechanisms that affect quantity and price. What mechanisms we will include depends on how widely we draw the boundaries of the system to be examined.

For example, we may include (1) a weather mechanism that determines the amount of rainfall; (2) a productivity mechanism that determines the yield of strawberries per acre; (3) a supply mechanism that determines the acreage sowed in strawberries; and (4) a demand mechanism that determines the quantity of strawberries purchased. In this formulation, each mechanism, which might be represented by an equation, determines the value of a particular variable as a function of some other variables (not specified in the account above). The variable whose value is so determined (dependent variable) may be called the *effect* of the working of that particular mechanism, while the values of other variables entering into the mechanism (independent variables) are the *causes* of that effect.

In the example before us, we might write:

$$R = r \tag{1}$$

$$Y = f_1(R,\ F, T) \tag{2}$$

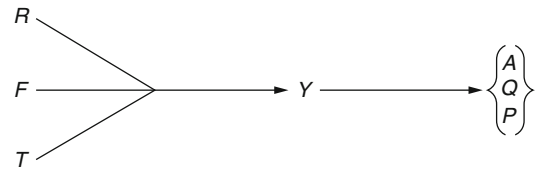
$$A = f_2(p) \tag{3}$$

$$Q = YA \tag{D}$$

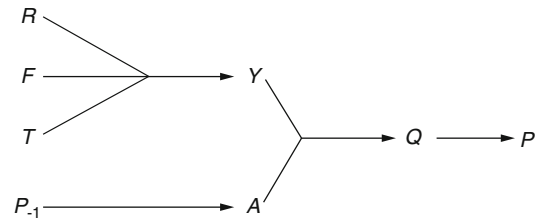
$$p = f_3(Q) \tag{4}$$

Here R is rainfall, and r a positive constant; Y is the yield per acre, F the amount of fertilization, and T the amount of tillage; A is the acreage sowed, and p the market price; and Q is the total yield. Equation D represents a definition, not a separate mechanism. In Eq. 4, p is taken as the dependent variable, since Q is assumed already to be determined by Eqs. 2 and 3 (cobweb assumption) (Fig. 1).

The system of equations defines a *causal ordering* among the variables. The value of R is determined exogenously, as are the values of F and T. That is to say, they are determined in some larger system of which the mechanisms described in the equations are only a subset. The value of Y follows from those of R, F, and T. The values of A, Q, and p are determined simultaneously. Thus, the equations determine a partial ordering:



Causality in Economic Nodels, Fig. 1



Causality in Economic Nodels, Fig. 2

Notice that the asymmetry that underlies this ordering cannot be interpreted as the asymmetry of logical implication, for from 'A implies B' we can infer that 'not-B implies not-A', while from, 'Heavy rainfall causes the yield to be large' we cannot conclude that 'A small yield causes a scanty rainfall.' The most accurate mode of expression is: 'The amount of rainfall determines (causes) the amount of yield' – large or small in both cases. The asymmetry reflects a distinction between exogeneity and endogeneity of variables, based, in turn, upon controllability (in the case of variables that can be manipulated directly), or time precedence. Thus R is exogenous to mechanism Eq. 2 on the assumption that the weather is unaffected by changes in the yield of strawberries. (That this is an empirical assumption is clear from the fact that widespread cultivation *can* cause changes in climate) (Fig. 2).

If we wish to remove the ambiguity from the causal relations among A, Q, and p, we may

assume (as in the classical cobweb theory) a time lag, replacing p in Eq. 3 by the exogenous and predetermined variable, p_{-1}. Then the causal ordering becomes:

Now it is clear why we took p as the dependent variable in Eq. 4. Introducing the time lag requires making an empirical assumption – specifically an assumption about how farmers form expectations about future prices. In a rational expectations model, for example, this lag would not be admissible.

Formalization

To formalize these ideas, consider a system of n simultaneous linear equations in n variables (*linear structure*). We assume that each equation represents a mechanism. In some linear structures, certain subsets of equations can be solved independently of the remaining equations (*self-contained subsets*). Consider the *minimal self-contained subsets* of a system (those that do not themselves contain smaller self-contained subsets). With each such subset, associate the variables that can be evaluated from the subset alone (endogenous variables), and call them *variables of order zero*. Next, substitute the values of these variables in the remaining equations of the system, and repeat the whole process, obtaining the variables of order one, two, and so on, and the corresponding minimal self-contained subsets of equations. If a variable of some order occurs with non-zero coefficient in an equation belonging to a subsystem of higher order, then that variable is one of the causes of the values of the endogenous variables of the latter set.

Thus, in our original example, Eq. 1 is the minimal self-contained subset of zero order, and R is its variable; Eq. 2 is the minimal subset of first order, and Y its variable; while Eqs. 3, D, and 4 constitute the minimal self-contained subset of second order, and A, Q, and p are their variables. The exogenous variables, F and T, can be regarded as parameters of the system, or equations parallel to Eq. 1 can be added for them, so that each belongs to a separate minimal self-contained subset of order zero.

Identifiability of Causal Ordering

The causal ordering among variables in a linear self-contained structure depends on which variables appear with non-zero coefficients in which equations. Consider a set of observations of the variables satisfying the equations of such a structure. Clearly these observations will also satisfy a new structure made up of equations that are arbitrary linear combinations of equations drawn from the original set. But different combinations of variables will generally appear in the equations of the new structure than appeared in the equations of the original structure. Taking these linear combinations 'blends' the separate mechanisms represented by the original equations. Hence, the causal ordering is not preserved under such a transformation, although the same empirical observations are compatible with both sets of equations.

From this consideration it follows that causal ordering cannot be inferred from simultaneous observations, no matter how numerous, of the variables of a structure. Additional assumptions must be made to identify a unique structure from the observations. The *identification problem* of econometrics is the problem of finding a sufficient number of prior assumptions to determine a unique set of equations, each corresponding to a mechanism, that fits the observations. The equations thus determined are usually called *structural equations*, while algebraically equivalent equations derived from them by linear combination are called *reduced form* equations.

The assumptions needed to identify structural equations may be derived from prior knowledge about mechanisms (e.g., our knowledge that the weather affects crops, but crops do not usually affect the weather). Where experimentation is possible, holding particular variables constant while varying others, experimental findings are a powerful source of empirically valid identifying assumptions. Sometimes, there is prior knowledge, also, that particular mechanisms are independent of each other (that farmers make their decisions independently of consumers, and vice versa). Whatever their source, the

identifying assumptions are genuine empirical assertions, and cannot be made arbitrarily or for reasons of statistical convenience if the correct causal inferences are to be drawn. So-called 'spurious' correlation is best interpreted as a relation between variables that does not have causal force because it was estimated from equations that did not correspond to independent mechanisms.

See Also

► Simultaneous Equations Models

Bibliography

Goldberger, A.S., and O.D. Duncan. 1973. *Structural equation models in the social sciences*. New York: Academic.
Hood, W.C., and T.C. Koopmans. 1953. *Studies in econometric method*. New York: Wiley.
Simon, H.A. 1977. Causes and possible worlds. Section 2. In *Models of discovery*, ed. H.A. Simon. Dordrecht: D. Reidel.

Causality in Economics and Econometrics

Kevin D. Hoover

Abstract

Economics was conceived as early as the classical period as a science of causes. The philosopher–economists David Hume and J. S. Mill developed the conceptions of causality that remain implicit in economics today. This article traces the history of causality in economics and econometrics, showing that different approaches can be classified on two dimensions: process versus structural approaches, and a priori versus inferential approaches. The variety of modern approaches to causal inference is explained and related to this classification. Causality is also examined in relationship to exogeneity and identification.

Keywords

Aristotle; Causal inference; Causality in economics and econometrics; Correlation; Cowles Commission; Econometrics; Edgeworth, F. Y.; Endogeneity and exogeneity; Granger–Sims causality; Graph theory; Hume, D.; Identification; Index numbers; Induction; Instrumental variables; Jevons, W. S.; Microfoundations; Mill, J. S.; Mises, L. von; Natural experiments; Observational equivalence; Process analysis; Quetelet, A.; Rational expectations; Regression; Robbins, L. C.; Simon, H.; Smith, A.; Statistical inference; Structural vector autoregressions; Tinbergen, J.; Vector autoregressions; Wold, H. O. A

JEL Classifications

B4

Philosophers of Economics and Causality

The full title of Adam Smith's great foundational work, *An Inquiry into the Nature and Causes of the Wealth of Nations* (1776), illustrates the centrality of causality to economics. The connection between causality and economics predates Smith. Starting with Aristotle, the great economists are frequently also the great philosophers of causality. Aristotle's contributions to economics are found principally in the *Topics*, the *Politics*, and the *Nicomachean Ethics*, while he lays out his famous four causes (material, formal, final and efficient) in the *Physics*. Material and formal causes are among the concerns of economic ontology, a subject addressed by philosophers of economics (see, for example, Mäki 2001) albeit rarely by practicing economists. Sometimes, as for example in Karl Marx's grand theory of capitalist development, economists have appealed to final causes or teleological explanation (for a defence, see Cohen 1978; for a general discussion, see Kincaid 1996). But, for the most part, taking physical sciences as a model, economics deals with efficient causes. What is it that makes things happen? What

explains change? (See Bunge 1963, for a broad account of the history and philosophy of causal analysis.)

The greatest of the philosopher/economists, David Hume, set the tone for much of the later development of causality in economics. On the one hand, economists inherited from Hume the sense that practical economics was essentially a causal science. In 'On Interest', Hume (1742, p. 304) writes:

> it is of consequence to know the principle whence any phenomenon arises, and to distinguish between a cause and a concomitant effect. Besides that the speculation is curious, it may frequently be of use in the conduct of public affairs. At least, it must be owned, that nothing can be of more use than to improve, by practice, the method of reasoning on these subjects, which of all others are the most important; though they are commonly treated in the loosest and most careless manner.

On the other hand, Hume doubted whether we could ever know the essential nature of causation 'in the objects' (Hume 1739, p. 165). Coupled with a formidable critique of inductive inference more generally, Hume's scepticism has contributed to a wariness about causal analysis in many sciences, including economics (1739, 1777). The tension between the epistemological status of causal relations and their role in practical policy runs through the history of economic analysis since Hume.

History

Hume's Foundational Analysis

Although Hume's dominant concerns are moral, historical, political, and social (including economic), physical illustrations serve as his paradigm causal relationships. *A* (say, a billiard ball) strikes *B* (another ball) and causes it to move. Any analysis must address two key features of causality: first, causes are asymmetrical (in general, if *A* causes *B*, *B* does not cause *A*). Hume sees temporal succession (the movement of *A* precedes the movement of *B*) as accounting for asymmetry. Second, causes are effective. A cause must be distinguished from an accidental correlation and must bring about its effect. Hume sees spatial contiguity (the balls touch) and necessary connection (the movement of *B* follows of necessity from the movement of *A*) as distinguishing causes from accidents and establishing their effectiveness.

Hume was famously sceptical of any idea that could not be traced either to logical or mathematical deduction or to direct sense experience. He asks, whence comes the idea of the necessary connection of cause and effect? It cannot be deduced from first principles. So, he argues that our idea of necessary connection, which he concedes is the most characteristic element of causality, can arise only from our experience of the constant conjunction of particular temporal sequences. But this then implies that causality stands on a very weak foundation. For one corollary of Hume's belief that all ideas are based either in logic or sense experience was that we do not have any secure warrant for inductive inference. Neither logic nor experience (unless we beg the question by implicitly assuming the truth of induction) gives us secure grounds from observing instances to inferring a general rule. Therefore, what we regard as necessary connection in causal inference is really more of habit of mind without clear warrant. Causes may be necessarily connected to effects; but, for Hume, we shall never know in what that necessary connection consists.

While later philosophers have differed with Hume on the analysis of causality, his views were instrumental in setting the agenda, not only for philosophical discussions, but for practical causal analysis as well.

The 19th Century: Logic and Statistics

Even more influential than Hume in shaping economics, John Stuart Mill, another philosopher/economist, was less sceptical about causal inference in general, but more sceptical about its application to economics. In his *System of Logic* (1851), Mill advanced his famous canons of induction: the methods of (*a*) agreement, (*b*) difference, (*c*) joint (or double) agreement and difference, (*d*) residues, and (*e*) concomitant variations. For example, according to the method of difference, if we have two sets of

circumstances, one in which a phenomenon occurs and one in which it does not, and the circumstances agree in all but one respect, that respect is the cause of the phenomenon. Mill's canons are essentially abstractions from the manner in which causes are inferred in controlled experiments. As such, Mill doubted that the canons could be easily applied to social or economic situations, in which a wide variety of uncontrolled factors are obviously relevant. Mill argued that economics was what Daniel M. Hausman (1992) has called an 'inexact and separate science', whose general principles were essentially known a priori and which held only subject to *ceteris paribus* clauses. Mill's apriorism proved to be hugely influential in later economics. Lionel Robbins (1935) expressed considerable scepticism about the place of empirical studies within economic science. Some Austrian economists, such as Ludwig von Mises (1966), went so far as to deny that economics could be an empirical discipline at all. Mill's apriorism also influenced those economists who see economic theory as similar to physical theory as a domain of universal laws.

Other 19th-century economists were less sceptical about the application of causal reasoning to economic data. For instance, W. Stanley Jevons (1863) pioneered the construction of index numbers as the core element of an attempt to prove the causal connection between inflation and the increase in worldwide gold stocks after 1849. Jevons's investigation can be interpreted as an application of Mill's method of residues (see Hoover and Dowell 2001). He saw the various idiosyncratic relative price movements, owing to supply and demand for particular commodities, as cancelling out to leave the common factor that could only be the effect of changes in the money stock.

The 19th century witnessed extensive development in the theory and practice of statistics (Stigler 1986). Inference based on statistical distributions and correlation measures was closely connected to causality. Adolphe Quetelet envisaged the inferential problem in statistics as one of distinguishing among constant, variable, and accidental causes (Stigler 1999, p. 52). The economist Francis Ysidro Edgeworth pioneered tests of statistical significance (in fact Edgeworth may have been the first to use this phrase). He glossed the finding of a statistically significant result as one that 'comes by cause' (Edgeworth 1885, pp. 187–8).

The 20th Century: Causality and Identification

Further developments of statistical techniques, such as multiple correlation and regression, in the 20th century were frequently associated with causal inference. It was fairly quickly understood that, unlike correlation, regression has a natural direction: the regression of Y on X does not produce coefficient estimates that are the algebraic inverse of those from the regression of X on Y. The direction of regression should respect the direction of causation.

By the early 20th century, however, the dominant vision of economics was one in which prices and quantities are determined simultaneously. This is as much true for Alfred Marshall (1930), who is often described (not perfectly accurately) as an advocate of partial equilibrium analysis, as it is for Léon Walras (1954), the principal font of modern general equilibrium analysis. Simultaneity does not necessarily rule out causal order, though it does complicate causal inference. Although regressions may have a natural causal direction, there is nothing in the data on their own that reveal which direction is the correct one – each is an equally eligible rescaling of a symmetrical and non-causal correlation. This is a problem of observational equivalence. And it is the obverse side of the now familiar problem of econometric identification: in this case, how can we distinguish a supply curve from a demand curve? The problem of identification was pursued throughout most of the first half of the 20th century until the fairly complete treatment by the Cowles Commission at mid-century (Koopmans 1950; Hood and Koopmans 1953; see Morgan 1990, for a thorough treatment of the history of the identification problem).

The standard solution to the identification problem is to look for additional causal determinants that discriminate between otherwise simultaneous relationships. Both the supply of milk and

demand for milk depend on the price of milk. If, however, the supply also depends on the price of alfalfa used to feed the cows and the demand also on the daily high temperature (which affects the demand for milk to make ice cream), then supply and demand curves can be identified separately. Identification can be viewed through the glasses of simultaneous equations, pushing causality into the background, or it can be viewed as a problem in causal articulation. In the first case, economists frequently use the language of exogenous variables (the price of alfalfa, the temperature) and endogenous variables (the price and quantity of milk). Exogenous variables can also be regarded as the causes of the endogenous variables. From the 1920s to the 1950s, different economists placed different emphasis on the causal aspects of identification (Morgan 1990) and the various papers reprinted in Hendry and Morgan (1995).

Modern econometrics can be dated from the development of structural econometric models following the pioneering work in the 1930s of Jan Tinbergen, the conceptual foundations of probabilistic econometrics in Trgyve Haavelmo's (1944) 'Probability approach to econometrics', and the technical elaboration of the identification problem in the two Cowles Commission volumes. Structural models did not in themselves necessarily favor the language of identification over the language of causality. Indeed, in Tinbergen's (1951) textbook, dynamic, structural models are explicated with a diagram that uses arrows to indicate causal connections among time-dated variables. Nevertheless, after the econometric work of the Cowles Commission, two approaches can be clearly distinguished.

One approach, associated with Hermann Wold and known as *process analysis,* emphasized the asymmetry of causality, typically grounded it in Hume's criterion of temporal precedence (Morgan 1991). Wold's process analysis belongs to the time-series tradition that ultimately produced Granger causality and the vector autoregression (see section "Alternative Approaches to Causality in Economics").

The other approach, associated with the Cowles Commission, related causality to the invariance properties of the structural econometric model. This approach emphasized the distinction between endogenous and exogenous variables and the identification and estimation of structural parameters. Implicitly, structural modellers accepted Mill's a priori approach to economics. While they differed from Mill in their willingness to conduct empirical investigations, the selection of exogenous (or *instrumental*) variables was seen to be the province of a priori economic theory – a maintained assumption rather than something to be learned from data itself.

In his contribution to one of the Cowles Commission volumes, Herbert Simon (1953) showed that causality could be defined in a structural econometric model, not only between exogenous and endogenous variables, but also among the endogenous variables themselves. And he showed that the conditions for a well-defined causal order are equivalent to the well-known conditions for identification. Despite the equivalence, with the demise of process analysis and the ascendancy of structural econometrics – aided indirectly perhaps by a revival of Humean causal scepticism among the logical-positivist philosophers of science – causal language in economics virtually collapsed between 1950 and about 1990 (Hoover 2004).

Alternative Approaches to Causality in Economics

Different approaches to causality can be classified along two lines as shown in Fig. 1. One the one hand, approaches may emphasize structure or process. On the other hand, approaches may rely on a priori identifying assumptions or they may seek to infer causes from data. The upper left cell, the a priori structural approach, represented by the Cowles Commission, dominated economics for most of the postwar period. But since we already discussed it at some length in section "History", and since it was largely responsible for turning the economics profession away from *explicit* causal analysis, we add nothing more about it here and instead turn to the other cells in Fig. 1.

Causality in Economics and Econometrics, Fig. 1 Classification of approaches to causality in economics

	Structural	*Process*
A Priori	Cowles Commission: Koopmans (1953); Hood and Koopmans (1953)	Zellner (1979)
Inferential	Simon (1953) Hoover (1990; 2001) Favero and Hendry (1992) Natural experiments: Angrist and Krueger (1999; 2001)	Granger (1969) Vector autoregressions: Sims (1980)

The Inferential Structural Approach

The most important of the inferential structural approaches is due to Simon (1953). Simon eschews temporal order as a basis for causal asymmetry and, instead, looks to recursive structure. As we observed in section "History", Simon's account is closely related to the Cowles Commission's structural approach. Consider the bivariate system:

$$Y_t = \theta X_t + \varepsilon_{1t}, \tag{1}$$

$$X_t = \varepsilon_{2t}, \tag{2}$$

where the random error terms ε_{it} are independent, identically distributed and θ is a parameter. Simon says that X_t causes Y_t, because X_t is recursively ordered ahead of Y_t. One knows all about X_t without knowing about Y_t, but one must know the value of X_t to determine the value of Y_t. Equations (1) and (2) also appear to show that any intervention in (2), say a change in the variance of ε_{2t}, would transmit to (1); while any intervention in (1), say a change in θ or the variance of ε_{1t}, would not transmit to (2). Apparently, X_t could then be used to control Y_t.

Unfortunately, merely being able to write an accurate description of the two variables in the form of (1) and (2) does not guarantee either the apparent asymmetry of information or control. The same data can be repackaged into a statistically identical form with an apparently different causal order. For example, consider the following related system:

$$Y_t = \omega_{1t}, \tag{3}$$

$$X_t = \delta Y_t + \omega_{2t}, \tag{4}$$

where $\delta = \dfrac{\theta \mathrm{var}(\varepsilon_2)}{\theta^2 \mathrm{var}(\varepsilon_2) + \mathrm{var}(\varepsilon_1)}$, $\omega_{1t} = \varepsilon_{1t} + \theta\varepsilon_{2t}$, and $\omega_{2t} = (1 - \delta\theta)\,\varepsilon_{2t} - \delta\varepsilon_{1t}$.

Equations (3) and (4) are derived from Eqs. (1) and (2). The details of the algebra are not important. Essentially, (3) and (4) are linear combinations of (1) and (2) with multiplicative factors carefully chosen, so that the error terms ω_{1t} and ω_{2t} are uncorrelated. Such linear combinations preserve the values of X_t and Y_t and their statistical likelihood (that is, the two systems of equations have the same reduced form) and, so, describe the data equally well. Equations (3) and (4) have a form analogous to (1) and (2); but, on Simon's criterion, it appears that Y_t causes X_t on Simon's criterion. While it looks like the key parameters for (3) and (4) are derived from those of (1) and (2), we could have taken (3) and (4) as the starting point and derived (1) and (2) symmetrically. What we would like to do is to replace the equal signs with arrows that show that the causal direction runs from the right-hand to the left-hand sides in the regression equations in one of the systems, but not in the other. Unfortunately, there is no way to do this, no choosing between the systems, on the basis of a single set of data by itself. This is the problem of observational equivalence again.

The a priori approach of the Cowles Commission relies on economic theory to provide appropriate identifying assumptions to resolve the observational equivalence. Christopher Sims

(1980) attacked the typical application of the Cowles Commission's approach to structural macroeconometric models as relying on 'incredible' identifying assumptions: economic theory was simply not informative enough to do the job. But Simon, who was otherwise supportive of the conception of causality in the Cowles Commission, took a different tack.

Simon sees the problem as choosing between two alternative sets of parameters: which set contains the structural parameters, $\{\theta$ and the variances of the $\varepsilon_{it}\}$ or $\{\delta$ and the variances of the $\omega_{it}\}$? Simon suggested that experiments – either controlled or natural – could help to decide. If, for example, an experiment could alter the conditional distribution of X_t without altering the marginal distribution of Y_t, then it must be that Y_t causes X_t, because this would be possible only if a structure like (3) and (4) characterized the data. If it did, a change in the conditional distribution would involve either δ or the variance of ω_{2t}, neither of which would affect the variance of ω_{1t}. In contrast, if (1) and (2) truly characterized the causal structure of the data, a change to the conditional distribution of X_t would, in fact, involve a change to the variance of ε_{2t}, which, according to the equivalences above, would alter either δ or the variance of ω_{2t}. Similar relationships of stability and instability in the face of changes to the marginal distribution can also be demonstrated (Hoover 2001, ch. 7). The appeal to experimental evidence is what marks Simon's approach out as inferential rather than a priori.

Hoover (1990, 2001) generalizes Simon's approach to the type of nonlinear systems of equations found in modern rational-expectations models. He shows that Simon's idea of natural experiments can be operationalized by coordinating historical, institutional, or other non-statistical information with information from structural break tests on what, in effect, amounts to the four regressions corresponding to (1), (2), (3), and (4) above generalized to include lagged dynamics. With allowances for complications introduced by rational expectations, the key idea is that, in the true causal order, interventions that alter the parameters governing the true marginal distribution do not transmit forward to the conditional distribution (characterized by (1) or (4)) nor do interventions in the true conditional distribution transmit backward to the marginal distribution (characterized by (2) or (3)). Since the true structural parameters are not known a priori, non-statistical information is important in identifying an intervention as belonging to the process governing one variable or another.

Although avoiding the term 'causality', Favero and Hendry's (1992) analysis of the Lucas critique in terms of 'super-exogeneity' is also a variant on Simon's causal analysis (Ericsson and Irons 1995; Hoover 2001, ch. 7). Super-exogeneity is essentially an invariance concept (Engle et al. 1983). Favero and Hendry find evidence against the Lucas critique (non-invariance in the face of changes in policy regime) in the super-exogeneity of conditional probability distributions in the face of structural breaks in marginal distributions – the same sort of evidence that Hoover cites as helping to identify causal direction.

The recent revival of causal analysis in microeconomics in the guise of 'natural experiments', although apparently developed independently of Simon, nonetheless proceeds in much the same spirit as Hoover's version of Simon's approach (Angrist and Krueger 1999, 2001). This literature typically employs the language of instrumental variables. A natural experiment is a change in a policy or a relevant environmental factor that can be identified non-statistically. Packaged as an econometric instrument, the experiment can be used – in much the same way that variations in alfalfa prices and temperature were used in the example in section "History" – to identify the underlying relationships and to measure the causally relevant parameters.

While the development of structural approaches in econometrics has largely been independent, there is some cross-fertilization between economists and philosophers (for example, Simon and Rescher 1966); and recently philosophers of causality have looked to economics for inspiration and examples (for example, Cartwright 1989; Woodward 2003).

The Inferential Process Approach

Perhaps the most influential explicit approach to causality in economics is due to Clive

C

W. J. Granger (1969). Granger causality is an inferential approach, in that it is data-based without direct reference to background economic theory; and it is a process approach, in that it was developed to apply to dynamic time-series models (see Granger–Sims causality in this dictionary for technical details). Granger–Sims causality is an example of the modern probabilistic approach to causality, which is a natural successor to Hume (for example, Suppes 1970). Where Hume required constant conjunction of cause and effect, probabilistic approaches are content to identify cause with a factor that raises the probability of the effect: A causes B if $\mathrm{P}(B|A) > \mathrm{P}(B)$, where the vertical '|' indicates 'conditional on'. The asymmetry of causality is secured by requiring the cause (A) to occur before the effect (B) (but the probability criterion is not enough on its own to produce asymmetry since $P(B|A) > P(B)$ implies $\mathrm{P}(A|B) > \mathrm{P}(A)$).

Granger's (1980) definition is more explicit about temporal dynamics than is the generic probabilistic account, and it is cast in terms of the incremental predictability of one variable conditional on another:

X_t Granger-causes Y_{t+1} if $\mathrm{P}(Y_{t+1}|$ all information dated t and earlier$) \neq \mathrm{P}(Y_{t+1}|$ all information dated t and earlier omitting information about X).

This definition is conceptual, as it is impracticable to condition on *all* past information.

In practice, Granger causality tests are typically implemented through bivariate regressions. As an illustration, consider the regression equations:

$$Y_t = \Pi_{11} Y_{t-1} + \Pi_{12} X_{t-1} + \upsilon_{1t}, \qquad (5)$$

$$X_t = \Pi_{21} Y_{t-1} + \Pi_{22} X_{t-1} + \upsilon_{2t}, \qquad (6)$$

where the Π_{ij} are parameters, and the υ_{it} are random error terms. In practice, lag lengths may be larger than one, but far less than the infinity implicit in the general definition. X_t Granger-causes Y_{t+1} if $\Pi_{12} \neq 0$, and Y_t Granger-causes X_{t+1} if $\Pi_{21} \neq 0$.

Sims (1972) famously used Granger causality to demonstrate the causal priority of money over nominal income. Later, as part of a generalized critique of structural econometric models, Sims (1980) advocated vector autoregressions (VARs) – atheoretical time-series regressions analogous to Eqs. (1) and (2), but generally including more variables with lagged values of each appearing in each equation. In the VAR context, Granger causality generalizes to the multivariate case.

While Granger causality has something useful to say about incremental predictability, there is no close mapping between Granger causality and structural notions of causality on either the Cowles Commission's or Simon's accounts (Jacobs et al. 1979). Consider a structural model:

$$Y_t = \theta X_t + \beta_{11} Y_{t-1} + \beta_{12} X_{t-1} + \varepsilon_{1t}, \qquad (7)$$

$$X_t = \gamma Y_t + \beta_{21} Y_{t-1} + \beta_{22} X_{t-1} + \varepsilon_{2t}, \qquad (8)$$

where ε_{1t} and ε_{2t} are identically distributed, independent random errors and θ, γ, and the β_{ij}s are structural parameters. The independence of the parameters and the error terms implies that causality runs from the right-hand to the left-hand sides of each equation. Equations (5) and (6) can be seen as the reduced forms of (7) and (8).

We focus on X causing Y. X structurally causes Y if either θ or $\beta_{12} \neq 0$. And X Granger causes Y if $\Pi_{12} = \frac{\beta_{12} + \theta\beta_{22}}{1 - \theta\gamma} \neq 0$. Thus, if X Granger causes Y, then X structurally causes Y. Note, however, that this result is particular to the case in which (7) and (8) represents the universe, so that (5) and (6) represent the complete conditioning on past histories of relevant variables. If the universe is more complex and the estimated VAR does not capture the true reduced forms of the structural system, which in practice they may not, then the strong connection suggested here does not follow.

More interestingly, even if (5), (6), (7), and (8) are complete, structural causality does not necessarily imply Granger causality. Suppose that $\beta_{12} = \beta_{22} = 0$, but $\theta \neq 0$, then X structurally causes Y, but since $\Pi_{12} = 0$, X does not Granger cause Y.

Now suppose that X does not Granger cause Y. It does not necessarily follow that X does not structurally cause Y, since if θ, β_{12}, and $\beta_{22} \neq 0$, and $-\beta_{12}/\beta_{22} = \theta$, then it will still be true that

$\Pi_{12} = 0$. This may appear to be an odd special case, but in fact conditions such as $-\beta_{12}/\beta_{22} = \theta$ arise commonly in optimal control problems in economics.

A simple physical example makes it clear what is happening. Suppose that X measures the direction of the rudder on a ship and Y the direction of the ship. The ship is pummeled by heavy seas. If the helmsman is able to steer on a straight course, effectively moving the rudder to exactly cancel the shocks from the waves, the direction of the rudder (in ignorance of the true values of the shocks) will not predict the course of the ship. The rudder would be structurally effective in causing the ship to turn, but it would not Granger-cause the ship's course.

The a Priori Process Approach

The upper right-hand cell of Fig. 1 is represented by Arnold Zellner's (1979) account of causality (cf. Keuzenkamp 2000, ch. 4, s. 4). Zellner's notion of causality is borrowed from the philosopher Herbert Feigl (1953, p. 408), who defines causation '... in terms of predictability according to law (or more adequately, according to a set of laws)'. On the one hand, Zellner opposes Simon and sides with Granger: predictability is a central feature of causal attribution, which is why his is a process account. On the other hand, he opposes Granger and sides with Simon: an underlying structure (a set of laws) is a crucial presupposition of causal analysis, which is why his is an a priori account.

Much obviously depends on what a law is. Zellner's own view is that a law is a (probabilistic) description of a succession of states of the world that holds for many possible boundary conditions and covers many possible circumstances. He couches his position in an explicitly Bayesian theory of inference. Feigl identifies causality with lawlikeness or predictability. It is the fact that formulae fit previously unexamined cases, as well as examined ones, which constitutes their lawlikeness. This is close to Simon's invariance criterion (the true causal order is the one that is invariant under the right sort of intervention).

The central problem, then, is how to distinguish laws from false generalizations or accidental regularities – that is, how to distinguish conditional relations invariant to interventions from regularities that are either not invariant or are altogether adventitious. Zellner believes that a theory serves as the basis for discriminating between laws and casual generalizations. Although Zellner's approach permits us to learn some things from the data, in keeping with the spirit of Bayesian inference, it does so within a narrowly defined framework (cf. Savage's 1954, pp. 82–91, 'small world' assumption). Economic theory in Zellner's account restricts the scope of an investigation a priori.

Zellner objects to Granger causality for two reasons. First, it is not satisfactory to identify cause with temporal ordering, as temporal ordering is not the ordinary, scientific or philosophical foundation of the causal relationship. Second, Granger's approach is atheoretical. In order to implement it practically, an investigator must impose restrictions – limit the information set to a manageable number of variables, consider only a few moments of the probability distribution (in our exposition, just the mean), and so forth. For Zellner, if these restrictions cannot be explained theoretically, Granger's methods will discover only accidental regularities.

Zellner explicitly criticizes Granger for ignoring the need for theoretical basis for empirical investigation – implicitly focusing on only one side of a process in which theory informs empirics and empirics inform theory. He criticizes Simon for defining cause to be a formal property of a model (recursive order) without making essential reference to empirical reality. Zellner's criticism is, however, more aptly directed at the Cowles Commission's approach, since (as we saw in section "The Inferential Structural Approach") Simon distinguishes himself through tying causal order to empirical inference.

Structural Vector Autoregressions

Not all approaches to causality fall quite neatly into the cells of Fig. 1; or, more to the point, an approach that falls into one cell may morph into one that falls into another cell. The history of Sims's VAR program is an important case.

Sims (1980) advocated VARs as a reaction to the manner in which the Cowles Commission

programme, which identified structural models through a priori theory, had been implemented (see section "The Inferential Process Approach"). From a causal perspective, it was closely related to Granger's analysis. Starting with VAR such as Eqs (5) and (6), Sims wished to work out how various 'shocks' would affect the variables of the system. This is complicated by the fact that the error terms in (5) and (6), which might be taken to represent the shocks, are not in general independent, so that a shock to one is a shock to both, depending on how correlated they are. Sims's initial solution was to impose an arbitrary orthogonalization of the shocks (a Choleski decomposition). In effect, this meant transforming (5) and (6) into a system like (6) and (7) and setting either θ or γ to zero. This amounts to imposing a recursive order on X_t and Y_t, such that the covariance matrix of the error terms is diagonal (that is, ε_{1t} and ε_{2t} are uncorrelated). A shock to X can then be represented by a realization of ε_{1t} and a shock to Y by a realization of ε_{2t}.

Initially, Sims treated the choice of recursive order as a matter of indifference. Criticizing the VAR program from the point of view of structural models, Leamer (1985) and Cooley and LeRoy (1985) pointed out that the substantive results (for instance, impulse-response functions and innovation accounts) depend on which recursive order is chosen. Sims (1982, 1986) accepted the point and henceforth advocated Structural vector autoregressions (SVARs). SVARs can be identified through the contemporaneous causal order only. So, for example, to identify (5) and (6), it is enough to assume that either θ or γ in (7) or (8) is zero; one need not make any assumptions about the β_{ij}s. Ironically, since the initial impulse behind the VAR programme was to avoid theoretically tenuous identifying assumptions, the choice of restrictions on contemporaneous variables used to transform the VAR into the SVAR are typically only weakly supported by economic theory.

Nevertheless, the move from the VAR to the SVAR is a move from an inferential to an a priori approach. It is also a move from a fully non-structural, process approach to a partially structural approach, since the structure of the contemporaneous variables, though not of the lagged variables, is fully specified. The SVAR approach can, therefore, be seen as straddling the cells on the first line of Fig. 1.

The Graph-Theoretic Approach to Causal Inference

A final approach to causality in economics sometimes provides another example of an inferential structural approach, and sometimes straddles the cells on the second line of Table 1. Graph-theoretic approaches to causality were first developed outside of economics by computer scientists (for example, Pearl 2000) and philosophers (for example, Spirtes et al. 2000), but have recently been applied within economics (Swanson and Granger 1997; Akleman et al. 1999; Bessler and Lee 2002; Demiralp and Hoover 2003).

The key ideas of the graph-theoretic approach are simple (see Demiralp and Hoover 2003 or Hoover 2005 for a detailed discussion). Any structural model can be represented by a graph in which arrows indicate the causal order. Equations (1) and (2) are represented by $X \rightarrow Y$ and Eqs. (3) and (4) by $Y \rightarrow X$. More complicated structures can be represented by more complicated graphs. Simultaneity, for instance, can be represented by double-headed arrows. The graphs allow us easily to see the dependence or independence among variables. Pearl (2000) and Sprites et al. (2000) demonstrate the isomorphism between causal graphs and the independence relationships encoded in probability distributions. This isomorphism allows conclusions about probability distributions to be derived from theorems proven using the mathematical techniques of graph theory.

Many of the results of graph-theoretic analysis are straightforward. Suppose that $A \rightarrow B \rightarrow C$ (that is, A causes B causes C). A and C would be probabilistically dependent; but, conditional on B, they would be independent. Similarly for $A \leftarrow B \leftarrow C$. In each case, B is said to *screen* A from C. Suppose that $A \leftarrow B \leftarrow C$. Then, once again A and C would be dependent, but conditional on B, they would be independent. B is said to be the *common cause* of A and C. Now suppose that A and B are independent conditional on sets of variables that exclude C or its descendants, and

$A \rightarrow C \leftarrow B$, and none of the variables that cause A or B directly causes C. Then, conditional on C, A and B are dependent. C is called an *unshielded collider* on the path ACB. (A *shielded* collider would have a direct link between A and B.) These are the simplest relationships of probabilistic dependence and independence. More complex ones may also obtain in which A is independent of B only conditional on more than one other variable (say, C and D).

A number of causal search algorithms have been developed (Sprites et al. 2000). These start with information about correlations (or other tests of unconditional and conditional statistical independence) among variables. The most common of these, the PC algorithm, assumes that graphs are strictly recursive (known in the literature as *acyclical*) and starts with a graph in which all variables are causally connected with an unknown causal direction (represented by the headless arrow, '—'). It then tests for independence among pairs of variables, conditioning on sets of zero variables, then one, then two, and so forth until the set of variables is exhausted. Whenever it finds independence, it removes the causal connection between the variables in the graph. Once the graph is pared down as far as it can be, it considers triples of variables in which two are conditionally independent but are connected through a third. If conditioning on that third variable renders the variables conditionally dependent, then that variable is an unshielded collider and it is connected to the other two variables with causal arrows running toward it. After all the unshielded colliders have been identified, further logical analysis can be used to orient additional causal arrows. For example, we might reason as follows: suppose we have a triple $A \rightarrow C \rightarrow B$; unless the causal arrow runs away from C toward B, C would be identified as an unshielded collider; but C was not identified as an unshielded collider earlier in the search; therefore, the causal arrow must run away from C towards B, so that the graph becomes $A \rightarrow C \rightarrow B$.

Sometimes the data allow the complete orientation of a causal graph, but sometimes some causal connections are left undirected. In this case, the graph marks out an equivalence class, and the algorithm has identified 2! causal graphs consistent with the empirical probability distribution, where $n =$ the number of undirected causal connections.

While most applications of graph-theoretic methods assume that the true causal structures are recursive (that is, strictly acylical), economics frequently treats variables that are cyclical or simultaneously determined. Although the recursiveness assumption is restrictive, it is an assumption that is also frequently made in the SVAR literature. Some progress has been made in developing graph-theoretic search algorithms for cyclical or simultaneous causal systems (Pearl 2000, pp. 95–6, 142–3; Richardson 1996; Richardson and Spirtes 1999).

Swanson and Granger (1997) showed that estimates of the error terms of the VAR (the υ_{it} in Eqs (5) and (6)) can be treated as the original time-series variables purged of their dynamics. A causal order identified on such variables corresponds to the causal order necessary to convert a VAR into an SVAR. Demiralp and Hoover (2003) present Monte Carlo evidence that the PC algorithm is effective at selecting the true causal connections among variables and, when signal strengths are high enough, moderately effective at directing them correctly. Search algorithms can, therefore, reduce or even eliminate the need to appeal to a priori theory when identifying the causal order of an SVAR.

Where Simon's approach looked for relatively important interventions as a basis for causal inference to a structure, the graph-theoretic approach uses relatively routine random variations to identify patterns of conditional independence that map out causal structures. The two approaches are complementary: Simon's approach may be used to resolve the observational equivalence reflected in causal connections that remain undirected after the application of a causal search algorithm.

From Metaphysics to Econometric Practice

The analysis of causation was originally a branch of metaphysics. In moving from the scholastic to the practical, two deep divisions appeared among economists.

The first is the divide between those who believed that causality in economics could be characterized by relatively simple uniformities (the process approaches) and those who believed that it must be characterized by a rich understanding of the underlying mechanisms (the structural approaches). Economists debate the appropriate level at which to characterize either the uniformities or the mechanisms – individual or aggregate. But this debate over the microfoundations of macroeconomics is another story. The second divide is between those who believe that economic logic itself gives privileged insight into economic behaviour (a priori approaches) and those who believe that we must learn about economic behaviour principally through observation and induction (the inferential approaches).

These are old debates – unlikely to be resolved decisively to the satisfaction of all economists in the near future. How one aligns oneself in them largely determines which particular approaches to causality appear to be compelling in practical economic research.

See Also

- ▶ Endogeneity and Exogeneity
- ▶ Granger–Sims Causality
- ▶ Graph Theory
- ▶ Hume, David (1711–1776)
- ▶ Identification
- ▶ Mill, John Stuart (1806–1873)
- ▶ Simon, Herbert A. (1916–2001)
- ▶ Structural Vector Autoregressions
- ▶ Vector Autoregressions

Bibliography

Akleman, D.G., D.A. Bessler, and D.M. Burton. 1999. Modeling corn exports and exchange rates with directed graphs and statistical loss functions. In *Computation, causation, and discovery*, ed. C. Glymour and G.F. Cooper. Menlo Park/Cambridge, MA: MIT Press and American Association for Artificial Intelligence.

Angrist, J.D., and A.B. Krueger. 1999. Empirical strategies in labor economics. In *Handbook of labor economics*, ed. O. Ashenfelter and D. Card, Vol. 3-A. Amsterdam: North-Holland.

Angrist, J.D., and A.B. Krueger. 2001. Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives* 15(4): 69–85.

Bessler, D.A., and S. Lee. 2002. Money and prices: U.S. data 1869–1914 (a study with directed graphs). *Empirical Economics* 27: 427–446.

Bunge, M. 1963. *The place of the causal principle in modern science*. Cleveland: Meridian Books.

Cartwright, N. 1989. *Nature's capacities and their measurement*. Oxford: Clarendon Press.

Cohen, G.A. 1978. *Karl Marx's theory of history: A defense*. Princeton: Princeton University Press.

Cooley, T.F., and S.F. LeRoy. 1985. Atheoretical macroeconomics: A critique. *Journal of Monetary Economics* 16: 283–308.

Demiralp, S., and K.D. Hoover. 2003. Searching for the causal structure of a vector autoregression. *Oxford Bulletin of Economics and Statistics* 65(Supplement): 745–767.

Edgeworth, F.Y. 1885. Methods of statistics. In Royal Statistical Society of Britain, *Jubilee volume of the Statistical Society.* London: E. Stanford.

Engle, R.F., D.F. Hendry, and J.-F. Richard. 1983. Exogeneity. *Econometrica* 51: 277–304.

Ericsson, N., and J. Irons. 1995. The Lucas critique in practice: Theory without measurement. In *Macroeconometrics: Developments, tensions and prospects*, ed. K.D. Hoover. Boston: Kluwer.

Favero, C., and D.F. Hendry. 1992. Testing the Lucas critique: A review. *Econometric Reviews* 11: 265–306.

Feigl, H. 1953. Notes on causality. In *Readings in the philosophy of science*, ed. H. Feigl and M. Brodbeck. New York: Appleton-Century-Crofts.

Granger, C.W.J. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* 37: 424–438.

Granger, C.W.J. 1980. Testing for causality: A personal viewpoint. *Journal of Economic Dynamics and Control* 2: 329–352.

Haavelmo, T. 1944. The probability approach in econometrics. *Econometrica* 12(Supplement), iii–vi, 1–115.

Hausman, D.M. 1992. *The inexact and separate science of economics*. Cambridge: Cambridge University Press.

Hendry, D.F., and M.S. Morgan, ed. 1995. *The foundations of econometric analysis*. Cambridge: Cambridge University Press.

Hood, W., and T. Koopmans, ed. 1953. *Studies in econometric method*, Cowles commission monograph no. Vol. 14. New Haven: Yale University Press.

Hoover, K.D. 1990. The logic of causal inference: Econometrics and the conditional analysis of causality. *Economics and Philosophy* 6: 207–234.

Hoover, K.D. 2001. *Causality in macroeconomics*. Cambridge: Cambridge University Press.

Hoover, K.D. 2004. Lost causes. *Journal of the History of Economic Thought* 26: 149–164.
Hoover, K.D. 2005. Automatic inference of the contemporaneous causal order of a system of equations. *Econometric Theory* 21: 69–77.
Hoover, K.D., and M.E. Dowell. 2001. Measuring causes: Episodes in the quantitative assessment of the value of money. *History of Political Economy* 33: 137–161.
Hume, D. 1739. In *A treatise of human nature,* ed. L.A. Selby-Bigge. Oxford: Clarendon Press, 1888n.d.-a.
Hume, D. 1742. Of interest. In *Essays: Moral, political, and literary,* ed. E.F. Miller. Indianapolis: Liberty Classics, 1985n.d.-c.
Hume, D. 1777. An enquiry concerning human understanding. In *Enquiries concerning human understanding and concerning the principles of morals*, 2nd ed., ed. L.A. Selby-Bigge. Oxford: Clarendon Press, 1902n.d.-b.
Jacobs, R.L., E.E. Leamer, and M.P. Ward. 1979. Difficulties in testing for causation. *Economic Inquiry* 17: 401–413.
Jevons, W.S. 1863. A serious fall in the value of gold ascertained, and its social effects set forth. In *Investigations in currency and finance,* 1884, Reprinted. New York: Augustus M. Kelley, 1964.
Keuzenkamp, H.A. 2000. *Probability, econometrics, and truth*. Cambridge: Cambridge University Press.
Kincaid, H. 1996. *Philosophical foundations of the social sciences: Analyzing controversies in social research*. Cambridge: Cambridge University Press.
Koopmans, T., ed. 1950. *Statistical inference in dynamic economic models,* Cowles Commission Monograph No. 10. New York: Wiley.
Leamer, E.E. 1985. Vector autoregressions for causal inference. In *Understanding monetary regimes*, Carnegie-Rochester conference series on public policy, ed. K. Brunner and A.H. Meltzer, Vol. 22. Amsterdam: North-Holland.
Mäki, U. 2001. *The economic world view: Studies in the ontology of economics*. Cambridge: Cambridge University Press.
Marshall, A. 1930. *Principles of economics: An introductory volume*. 8th ed. London: Macmillan.
Mill, J.S. 1848. In *Principles of political economy with some of their applications to social philosophy*, ed. W.J. Ashley, 1909. London: Longman, Green.
Mill, J.S. 1851. *A system of logic, ratiocinative and deductive: Being a connected view of the principles of evidence and the methods of scientific investigation*. Vol. 1. 3rd ed. London: John W. Parker.
Morgan, M.S. 1991. The stamping out of process analysis in econometrics. In *Appraising economic theories: Studies in the methodology of research programs*, ed. N. De Marchi and M. Blaug. Aldershot: Edward Elgar.
Morgan, M.S. 1990. *The history of econometric ideas*. Cambridge: Cambridge University Press.
Pearl, J. 2000. *Causality: Models, reasoning, and inference*. Cambridge: Cambridge University Press.
Richardson, T. 1996. A discovery algorithm for directed cyclical graphs. In *Uncertainty in artificial intelligence: Proceedings of the twelfth congress*, ed. F. Jensen and E. Horwitz. San Francisco: Mogan Kaufman.
Richardson, T., and P. Spirtes. 1999. Automated discovery of linear feedback models. In *Computation, causation and discovery*, ed. C. Glymour and G.F. Cooper. Menlo Park: AAAI Press.
Robbins, L. 1935. *An essay on the nature and significance of economic science*. London: Macmillan.
Savage, L.J. 1954. *The foundations of statistics*, 1972. New York: Dover.
Simon, H.A. 1953. Causal order and identifiability. In Hood and Koopmans.
Simon, H.A., and N. Rescher. 1966. Causes and counterfactuals. *Philosophy of science* 33: 323–340.
Sims, C.A. 1972. Money, income and causality. *American Economic Review* 62: 540–552.
Sims, C.A. 1980. Macroeconomics and reality. *Econometrica* 48: 1–48.
Sims, C.A. 1982. Policy analysis with econometric models. *Brookings Papers on Economic Activity* 1982(1): 107–152.
Sims, C.A. 1986. Are forecasting models usable for policy analysis? *Federal Reserve Bank of Minneapolis Quarterly Review* 10(1): 2–15.
Spirtes, P., C. Glymour, and R. Scheines. 2000. *Causation, prediction, and search*. 2nd ed. Cambridge, MA: MIT Press.
Stigler, S.M. 1986. *The history of statistics: Measurement of uncertainty before 1900*. Cambridge, MA: Belknap Press.
Stigler, S.M. 1999. *Statistics on the table*. Cambridge, MA: Harvard University Press.
Suppes, P. 1970. *A probabilistic theory of causality. Acta Philosophica Fennica 24*. Amsterdam: North Holland.
Swanson, N.R., and C.W.J. Granger. 1997. Impulse response functions based on a causal approach to residual orthogonalization in vector autoregressions. *Journal of the American Statistical Association* 92: 357–367.
Tinbergen, J. 1951. *Econometrics*. New York: Blakiston Company.
von Mises, L. 1966. *Human action: A treatise on economics*. 3rd ed. Chicago: Henry Regnery.
Walras, L. 1954. *Elements of pure economics*. London: Allen and Unwin.
Woodward, J. 2003. *Making things happen: A theory of causal explanation*. Oxford: Oxford University Press.
Zellner, A.A. 1979. Causality and econometrics. In *Three aspects of policy making: Knowledge, data and institutions*, Carnegie-Rochester Conference Series on Public Policy, ed. K. Brunner and A.H. Meltzer, Vol. 10. Amsterdam: North-Holland.

Cazenove, John (1788–1879)

J. M. Pullen

Cazenove wrote nine books and pamphlets on political economy, dealing with a wide variety of theoretical concepts and practical issues. In addition, he made a valuable contribution to political economy as an editor of Richard Jones's *Literary Remains* (1859, p. xl), and T.R. Malthus's *Definitions in Political Economy* (1853). There is also strong evidence to suggest that he was the anonymous editor of the second edition of Malthus's *Principles of Political Economy* (Pullen 1978). He contributed a review to the *British Critic* ('Chalmers – On Political Economy', October 1832, a vigorous criticism of James Mill and Ricardo), and could possibly have contributed others (Gordon 1985, pp. 17–19).

Malthus had a high regard for Cazenove. He recommended to his publisher that the first edition of his *Principles* should be reviewed by Cazenove in the *Quarterly Review* (letter of 26 January 1821, in the archives of John Murray), and he nominated Cazenove for membership of the Political Economy Club, at its second meeting in 1821. J.L. Mallet recorded in his diary that, in the Club debates, 'on most occasions Ricardo and Mill led on one side, and Malthus and Cazenove on the other'. In a letter to Thomas Chalmers of 6 February 1833, Malthus described Cazenove as 'a particular friend' and as 'a very clever man, and good political economist'. When Malthus died in 1834, Cazenove applied (unsuccessfully) for his position as professor of history and political economy at the East India College (James 1969, pp. 355–6). But Cazenove's friendship with Malthus, and his agreement with some of Malthus's main doctrines, did not prevent him from criticizing Malthus on occasions and adopting an independent line (Pullen 1978, pp. 293–4).

Cazenove's omission from the first two editions of *Palgrave's Dictionary* and the absence of any entry under 'Cazenove' in the *Index of Economic Articles* up to 1979, indicate that his writings have so far attracted very little attention. But Gordon (1985) has shown that Cazenove does not deserve this neglect. His writings are a worthwhile contribution to political economy in their own right, and an important part of the anti-Ricardian tradition. In particular, Cazenove opposed Say's Law and recognized the possibility of a general glut. Like Malthus he emphasized the role played by effective demand, and denied that continued economic growth can be achieved merely through saving and capital accumulation. He stressed the idea – also put forward by Malthus, under the name of 'the doctrine of proportions' – that economic progress requires a balance between saving and consumption.

Cazenove's grandfather, David de Cazenove, and his father, James de Cazenove, were merchants of French Huguenot origin who migrated to London from Geneva in 1777 (*Burke's Landed Gentry*). Cazenove was born and died in London. He appears to have worked in his father's firm, Jas. Cazenove & Co., which he described (1861, pp. 42–3) as 'a large commercial firm' with 'some sixty or seventy foreign correspondents'. He retired from the business early – in 1832, at the age of 44, he described himself as 'late a continental merchant' – but his literary output continued until 1861. His father's brother, Phillip Cazenove, founded the present London firm of stockbrokers, Cazenove & Co. His brother, Philip Cazenove, was for many years the senior partner of Cazenove & Co., but there is no evidence that John Cazenove ever worked in that firm (information from Mr H. de L. Cazenove, of Cazenove & Co.) John Cazenove's son, John Gibson Cazenove, MA, DD (1822–96) of Brasenose College, Oxford, was Chancellor of Edinburgh Cathedral and author of theological works.

Selected Works

1820. *A Reply to Mr Say's letter to Mr Malthus on the subject of the stagnation of trade*. London:

J.M. Richardson. (The copy annotated by Ricardo is in Edinburgh University Library. See Ricardo, *Works and Correspondence*, ed. P. Sraffa, Cambridge, Cambridge University Press, Vol. X, 405–10.)
1822. *Considerations on the accumulation of capital and its effects on profits and on exchangeable value*. London: J.M. Richardson.
1829. *Questions respecting the national debt and taxation stated and answered*. London: J.M. Richardson. (The British Library copy contains MS alterations, presumably intended for a 2nd edition.)
1832a. *Outlines of political economy* . . .. London: Pelham Richardson.
1832b. The Evidence that WOULD have been given by Mr ——, late a continental merchant, before the Committee of Secrecy appointed to inquire into the expediency of renewing the Bank Charter. London: Pelham Richardson. (The British Library copy contains a MS note by Cazenove correcting an error on p. 15.)
1840. An elementary treatise on political economy . . .. London: A.H. Baily. 1847. The Money Crisis. London.
1861. Supplement to thoughts on a few subjects of political economy. London: Simpkin Marshall.
1859. Thoughts on a few subjects of political economy. London: Simpkin Marshall.

Bibliography

Gordon, B. 1985. John Cazenove (1788–1879): Critic of Ricardo, friend and editor of Malthus. Paper presented at the Third Conference of the History of Economic Thought Society of Australia held at La Trobe University.
James, P. 1979. *Population Malthus: His life and times*. London: Routledge & Kegan Paul.
Jones, R. 1859. *Literary remains, consisting of lectures and tracts on political economy, of the late Rev. Richard Jones*. Edited, with a Prefatory Notice, by the Rev. William Whewell, D.D., London: John Murray.
Malthus, T.R. 1853. *Definitions in political economy . . . A new edition, with a preface, notes, and supplementary remarks by John Cazenove*. London: Simpkin & Marshall.
Pullen, J.M. 1978. The editor of the second edition of T.R. Malthus' Principles of Political Economy. *History of Political Economy* 10(2): 286–297.

Censored Data Models

G. S. Maddala

The *censored* normal regression model considered by Tobin (1958), also commonly known as the 'tobit' model, is the following:

$$y_i^* = \beta x_i + u_i \quad u_i \sim \mathrm{IN}(0, \sigma^2)$$

The observed y_i are related to y_i^* according to the relationship

$$\begin{aligned} y_i &= y_i^* \quad \text{if } y_i^* > y_0 \\ &= y_0 \quad \text{otherwise} \end{aligned} \tag{1}$$

where y_0 is a prespecified constant (usually zero). They y_i^* could take values $<y_0$. The only thing is that they are not observed. Thus, y_i is set equal to y_0 because of *non-observability*. The values x_i are observed for all the observations. If *both* y_i and x_i are unobserved for $y_i \leq y_0$ then we have what is known as a *truncated* regression model.

The problem is essentially one of missing data. Data on y are missing for some observations. Hence, we have to ask why data are missing. In some cases this is owing to the design of the experiment, as in the case of the data from the negative income tax experiment. These data have been analysed by Hausman and Wise (1977), who consider a truncated regression model. In almost all other cases y_0 is the outcome of choices of individuals. In this case the model is incomplete unless the determinants of y_0 are studied.

Some Early Developments

The first application of the censored regression model (1) is that of Tobin (1958) who studied the expenditures on durable goods by 735 non-farm households. y_i^* is the ratio of total durable expenditures to disposable income and

$y_0 = 0$. However, y_i is not equal to zero here because of non-observability, but because of individuals' choices. Thus, the censored regression model (1) is inappropriate for this problem. In fact, the tobit model is inappropriate for almost all the applications in which it has been used (including that by Tobin).

The model by Cragg (1971) considers this as a sequential decision problem. For the case of demand for automobiles, the decisions are whether or not to buy a car and how much to spend if the decision to buy a car is made. In this model we have the latent variable:

$$I_i = x_i\delta_1 + \eta_{1i} \quad \eta_{1i} \sim \text{IN}(0,1) \tag{2}$$

The subscript i denotes the ith individual. We observe the dummy variable D_i which is defined as

$$\begin{aligned} D_i &= 1 \, if \, I_1 > 0 \qquad (\text{buyers}) \\ &= 0 \text{ otherwise} \quad (\text{non} - \text{buyers}) \end{aligned} \tag{3}$$

For those who purchased a car, Cragg specifies a log normal model. Thus, denoting expenditures by y_i, we have

$$\log y_i = x_i\delta_2 + \eta_{2i} \quad \eta_{2i} \sim \text{IN}\left(0, \sigma^2\right) \tag{4}$$

The equation is defined only for the individuals for which $D_i = 1$.

In practice, however, it is questionable whether individuals make their decisions this way. The decision of whether or not to buy a car and how much to spend if a car is bought are often joint decisions. One can formulate this model in terms of two latent variables. Though there are several variants of this that one can think of, one formulation is the following.

$y_1 =$ the cost of the car the individual wants to buy.

$y_2 =$ the maximum expenditure the individual can afford.

The actual expenditure y is given by

$$\begin{aligned} y &= y_i \quad \text{if} \quad y_1 \le y_2 \\ &= 0 \quad \text{if} \quad y_1 > y_2 \end{aligned} \tag{5}$$

We can, in fact, consider y_1 and y_2 both to be log normal. This model is discussed in Nelson (1977), though not with reference to the example of automobile expenditures.

It is tempting to use the simple tobit model (1) every time that one has a bunch of zero (or other limit) observations on y. However, this is inappropriate. For instance, if hours worked for a number of individuals in a sample are zero, it does not mean that one can apply the tobit model to explain hours worked. One has to construct a model where hours worked are zero because of some decisions about labour force participation, in terms of reservation and market wages, as done by Heckman (1974). Estimation of this model from censored as well as truncated samples is discussed in Wales and Woodland (1980).

Selection Models

In the estimation of censored regression models we often have to formulate the censoring function that incorporates individual decisions. This function is also called a *selection criterion*. Usually the selection criterion involves the choice variables and other explanatory variables. Thus, the model is formulated as:

$$y_1 = \beta_1 x_1 + u_1 \quad \text{Choice 1} \tag{6}$$

$$y_2 = \beta_2 x_2 + u_2 \quad \text{Choice 2} \tag{7}$$

and

$$I^* = \gamma_1 y_1 + \gamma_2 y_2 + \beta_3 x_3 + u \quad \text{Selection criterion} \tag{8}$$

The observed y is defined as

$$\begin{aligned} y &= y_1 \quad \text{if} \quad \text{I}^* > 0 \\ &= y_2 \quad \text{if} \quad I^* \le 0. \end{aligned}$$

Interest centres on the determinants of γ_1 and γ_2 (see Lee 1978; Willis and Rosen 1979). One can substitute y_1 and y_2 in (8) and get a reduced form for the selection criterion. In this approach interest mainly centres on the so-called 'selectivity bias' in the estimation of (6) and (7) by OLS.

Since both y_1 and y_2 are censored, we have to estimate the parameters in (6) and (7) by the use of ML methods. Heckman (1979) suggests a simple correction to the OLS, which involves the addition of an extra explanatory variable to each of (6) and (7) obtained from the estimation of the criterion function (8) in its reduced form. This criterion is based on the assumption of normality. Goldberger (1983) made some calculations with alternative error distributions and showed that this adjustment for selection bias is quite sensitive to departures from normality.

There have been two solutions to this problem. One is the extension of the analysis of selectivity to general error distributions. This is the approach considered in Lee (1982, 1983), a summary of which is also given in Maddala (1983, pp. 272–275) along with earlier suggestions by Olsen. The other alternative approach is to consider distribution-free estimates (see Cosslett 1984), though this methodology is in early stages of development. Thus, there are computationally feasible alternatives available to explore the selectivity problem without assuming normality and there are procedures available to test the assumption of normality as well (see Lee and Maddala 1985).

The 'Heckman correction' for selectivity bias is very popular, mainly because it is easy to apply. But for this same reason it has also been applied in cases where it is not applicable; such cases are cited in Maddala (1985).

Some Other Problems

Many of the problems connected with the estimation of the censored regression model, assuming parametric distributions, are discussed in Maddala (1983, Chaps. 6 and 9) and Amemiya (1984). For distribution-free methods one can refer to Miller and Halpern (1982), Cosslett (1984) and Powell (1984). It is now well known that the properties of the estimators change with the violation of some basic assumptions. For instance, heteroskedasticity and errors in the dependent variable do not affect the consistency property of OLS estimators in the normal regression model. With the censored regression model, the ML estimators are no longer consistent under these assumptions. Stapleton and Young (1984) suggest that with errors in the dependent variable, the 'correct' ML estimation appears computationally difficult but find some alternative estimators promising.

There has been some progress made in the development of distribution free estimation and estimation when the standard assumptions are violated. For tests of some of the standard assumptions, see Lee and Maddala (1985).

See Also

- ▶ Latent Variables
- ▶ Limited Dependent Variables
- ▶ Logit, Probit and Tobit
- ▶ Selection Bias and Self-selection

Bibliography

Amemiya, T. 1984. Tobit models: a survey. *Journal of Econometrics* 24: 3–61.

Cosslett, S.R. 1984. Distribution-free estimator of a regression model with sample selectivity. Discussion Paper, CEDS, University of Florida.

Cragg, J.G. 1971. Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica* 39: 829–844.

Goldberger, A.S. 1983. Abnormal selection bias. In *Studies in Econometrics, Time-Series and Multivariate Analysis*, ed. S. Karlin, T. Amemiya, and L.A. Goodman. New York: Academic.

Hausman, J.A., and D.A. Wise. 1977. Social experimentation, truncated distributions, and efficient estimation. *Econometrica* 45: 919–938.

Heckman, J.J. 1974. Shadow prices, market wages, and labor supply. *Econometrica* 42: 679–693.

———. 1979. Sample selection bias as a specification error. *Econometrica* 47: 153–161.

Lee, L.F. 1978. Unionism and wage rates: a simultaneous equations mode with qualitative and limited dependent variables. *International Economic Review* 19: 415–433.

———. 1982. Some approaches to the correction of selectivity bias. *Review of Economic Studies* 49: 355–372.

———. 1983. Generalized econometric models with selectivity. *Econometrica* 51: 507–512.

Lee, L.F., and G.S. Maddala. 1985. The common structure of tests for selectivity bias, serial correlation, heteroscedasticity, and non-normality in the tobit model. *International Economic Review* 26: 1–20.

Maddala, G.S. 1983. *Limited Dependent and Qualitative Variables in Econometrics*. New York: Cambridge University Press.
———. 1985. A survey of the literature on selectivity bias as it pertains to health-care markets. In *Advances in Health Economics*, ed. R.M. Scheffler, and L.F. Rossiter. Greenwich: JAI Press.
Miller, R., and J. Halpern. 1982. Regression with censored data. *Biometrika* 69: 521–531.
Nelson, F.D. 1977. Censored regression models with unobserved stochastic censoring thresholds. *Journal of Econometrics* 6: 309–327.
Powell, J.L. 1984. Least absolute deviations estimation for censored regression models. *Journal of Econometrics* 25: 303–326.
Stapleton, D.C., and D.J. Young. 1984. Censored normal regression with measurement error on the dependent variable. *Econometrica* 52: 737–760.
Tobin, J. 1958. Estimation of relationships for limited dependent variables. *Econometrica* 26: 24–36.
Wales, T.J., and A.D. Woodland. 1980. Sample selectivity and the estimation of labor supply functions. *International Economic Review* 21: 437–468.
Willis, R.J., and S. Rosen. 1979. Education and self-selection. *Journal of Political Economy* 87: S5–S36.

Central Bank Communication

Michael Ehrmann and Marcel Fratzscher

Abstract

Since the early 1990s, communication has become a primary tool for monetary authorities in managing expectations, both of financial markets and of the wider public, and an important ingredient in making the central bank accountable. The rapidly growing literature on central bank communication clearly confirms the importance of communication in managing expectations, thereby enhancing the effectiveness of monetary policy. Yet there is a large degree of heterogeneity in communication practices across monetary authorities in the world, and there continues to be a lively and controversial debate about what constitutes an optimal communication strategy.

Keywords

Accountability; Central bank; Communication; Expectations; Financial markets; Independence; Inflation; Monetary policy; Objectives; Policy decisions; Predictability; Transparency

JEL Classifications

E52; E58

Central bank communication refers to the process by which monetary authorities convey information regarding their objectives, strategies and tools, as well as about their current assessment of the economic situation and the monetary stance. Such communication typically serves two purposes: making the central bank accountable, and enhancing the effectiveness of central bank policies.

Whereas only a few decades ago, transparency was usually seen as counterproductive to an effective conduct of monetary policy, it is nowadays considered best practice in central banking. One trigger for this has been the move to grant independence to more and more central banks, which in turn entails an obligation of those central banks to be more accountable.

Central banking laws often specify a number of obligations to ensure a central bank's accountability, which at the same time shape their communication policies. Testimonies to parliament and the requirement to deliver annual reports are examples. In several cases the relevant central banking acts also prescribe the targets for the monetary authority; their communication is therefore automatic, and not at the discretion of the central bank. At the same time, most central banks communicate substantially more often, and in much greater detail, than required by law.

As to the second purpose, it became increasingly clear throughout the 1990s that managing expectations is a central part of monetary policy, and that transparency is vital for that purpose. Given that communication is essential for accountability and transparency, central banks are now putting considerable effort into designing and conducting their communication policies.

Blinder et al. (2008), in their survey of the literature on central bank communication, derive the conditions under which central bank communication may matter for the conduct of monetary policy. A crucial issue in this regard is that a central bank usually has direct control only over a short-term interest rate, yet needs to influence interest rates at all maturities and to affect market expectations not just about current levels but about the future path of interest rates.

If the economic environment were constant, if the central bank was credibly committed to an unchanging policy rule, and if private agents had full information and rational expectations, the path of monetary policy could be inferred correctly from the central bank's observed actions (Woodford 2005). In reality, of course, none of these conditions are likely to hold. In particular, in a changing environment economic agents are subject to a continuous learning process. The central banks' views are also of interest to the public in a world of uncertainty and asymmetric information, especially given the complexity and extent of the information that feeds into monetary policy decisions, which often require judgment and the use of heuristics (Svensson 2003; King 2005).

The revolution in thinking and practice that has taken place over the recent decades can be exemplified with the case of the US Federal Reserve, which over the last 15 years has gone a long way towards greater transparency. For instance, it has started to issue statements instantaneously after each monetary policy decision, where the decision is not only announced, but also briefly explained, and to provide qualitative forward guidance about future monetary policy decisions. It has also expedited the release of the minutes of Federal Open Market Committee (FOMC) meetings, and it has increased the frequency and expanded the content of the publicly released forecasts for several economic variables made by FOMC members.

The Announcement of a Central Bank's Objective

If a central bank is granted independence from its government, it must be given a clearly defined mandate. This is generally done by defining central bank objectives, often in a quantified fashion. But even if a central bank is not given a quantitative objective, it often decides to provide its own quantification, or is required to do so. The potential effects of such a clarification and quantification are substantial. Not only do they make an independent central bank more (easily) accountable, since its actions can be assessed by cross-checking actual economic outcomes with those mandated; furthermore, the announcement of an objective, and in particular its quantification, provides a yardstick for the expectations of economic agents.

The available empirical evidence demonstrates that increased transparency about central banks' strategies, and in particular the announcement of an explicit inflation objective, has fostered central bank credibility as well as the predictability of the path of monetary policy. Moreover, the recent trend towards more transparent central banking practices has certainly played a considerable part in improving the short-term predictability of policy decisions by many central banks over recent decades (BIS 2004, pp. 73–80).

The announcement of central bank objectives may also have a direct bearing on economic outcomes. For instance, Benati (2008) finds that inflation persistence is considerably lower in countries with explicit inflation targets. Levin et al. (2004) furthermore show that in inflation-targeting countries, private sector inflation expectations are not correlated with lagged inflation, indicating that inflation expectations are better anchored. This evidence has been corroborated by Gürkaynak et al. (2009), who show that in some advanced inflation-targeting countries, long-term inflation expectations are less responsive to macroeconomic data releases than in the United States, where no explicit inflation objective has been announced.

However, the fact that the announcement of a central bank's objective has an effect on inflation expectations need not automatically imply that there will be an effect on the ultimate objective. This question needs to be settled empirically. The available evidence is rather inconclusive at this stage, with for example Kuttner and Posen (1999) arguing that inflation is lower in inflation-targeting countries, whereas Ball and Sheridan

(2005) cannot find any such evidence, given that also the countries in their control groups managed to achieve low inflation rates.

In sum, the empirical evidence suggests that the announcement of a central bank's objective is beneficial, since it eases the conduct of monetary policy through its effect on agents' expectations, and because it helps to achieve sound macroeconomic outcomes. At the same time, it does not seem to be the only means to achieve such outcomes.

The Announcement of Policy Decisions

It is common practice nowadays among central banks to inform the public about monetary policy decisions as soon as the decision has been taken. There is substantial evidence that this practice improves the markets' understanding of monetary policy considerably. For example, Lange et al. (2003) observe that the announcement of FOMC policy decisions since 1994 has enabled markets to improve their forecasts of monetary policy decisions. Furthermore, Demiralp and Jorda (2002) provide evidence that, by announcing changes to the intended federal funds rate in real time, it has been possible to move the federal funds rate with a smaller volume of open market operations, which indicates that the announcement of policy decisions can make policy implementation more efficient.

The Communication of the Current Assessment of the Economic Situation and the Monetary Policy Stance

By announcing an objective, and possibly releasing information about its monetary policy strategy, about the models used and about the variables considered in the economic analysis, a central bank aims to help the public better understand its broader framework and the way in which it reacts to different circumstances and contingencies. However, even if the broader framework is generally well understood, it will be impossible to communicate *ex ante* all contingencies in such a way that the public can always deduce perfectly the central bank's assessment, just by interpreting the incoming macroeconomic data. Regular communication of the central bank's assessment of the current economic situation and the monetary policy stance does therefore remain important. Accordingly, central banks often release statements that provide explanations for a given policy decision, publish inflation and growth forecasts, and deliver speeches in the inter-meeting period.

As argued above, central banks have recently achieved a high degree of short-term predictability. Accordingly, markets react predominantly not to the announcement of a decision, but to the communication surrounding it, such as any explanation of the underlying reasons and any forward-looking component. Gürkaynak et al. (2005) find that longer-term maturities in the yield curve react in particular to the forward-looking component of the communication.

In line with this, Reeves and Sawicki (2007) show that the collective forms of Bank of England communication have a rather strong market impact, such as the minutes of the committee meetings and the Inflation Report. Communication by individual committee members, such as speeches or interviews, has nonetheless also been shown to be important. Kohn and Sack (2004) find that the testimonies by the FOMC chairman have substantial effects on financial markets, throughout the entire maturity spectrum. Financial market responsiveness to committee members' speeches have been identified, for example in Ehrmann and Fratzscher (2007) for the Federal Reserve, the Bank of England and the European Central Bank (ECB).

Potential Risks

The evidence suggests that central bank communication is an important policy tool, with substantial effects on financial markets, and the potential to enhance the effectiveness of monetary policy making. However, as any effective tool, it needs to be properly utilized; otherwise, it can lead to undesired outcomes.

Communicating too frequently, or providing too much information, can be damaging if there is a limit to how much information can be digested

effectively (Kahneman 2003). A widespread example where central banks limit their transparency relates to the blackout periods, whereby committee members would typically not make public statements about monetary policy-related issues just before policy meetings. As shown by Ehrmann and Fratzscher (2009), there are good reasons for adhering to such a rule, because communication during the blackout period leads to excessive market volatility.

A debate has also centred around how central banks should communicate if they receive noisy signals themselves. In coordination games in the vein of Morris and Shin (2002), or in learning models like Dale et al. (2008), whether or not a central bank should communicate depends crucially on the relative precision of the central bank's and the private sector's information. However, it has been debated to what extent a case for limiting central bank communication can arise in these models. With regard to the coordination game literature, Svensson (2006) argues that central bank communication would need a much (and implausibly) lower signal-to-noise ratio than that of private information.

Clarity is essential for good communication. A possible risk to clarity can arise because monetary policy is typically set by committees rather than by single individuals. This can give rise to a 'cacophony problem' (Blinder 2004, ch. 2) if too many disparate voices on a topic confuse rather than clarify the message. Central banks take different approaches in that regard. Whereas FOMC members communicate their individual views to the public (Bernanke 2004), this is not so for the ECB, which now adheres to a one-voice policy (Jansen and De Haan 2006). Importantly, however, markets adapt to such differences in communication style, for example by reacting more strongly to statements by the chairperson of committees with dispersed communication, and more equally to statements by all committee members if these communicate in a collegial fashion (Ehrmann and Fratzscher 2007).

Open Issues

The recent research on central bank communication, surveyed in Blinder et al. (2008), has provided a large number of relevant insights. Central bank communication is an important policy tool, with substantial effects on financial markets, and the potential to enhance the efficiency of monetary policy making. However, what constitutes an optimal communication strategy remains an unsettled issue. There is a large diversity in the communication policies of central banks, because the design of communication policies must take into account the cultural and institutional environment in which a central bank operates. Accordingly, it is evident that 'one size does not fit all'.

Other issues and debates remain unresolved at the time of writing. For instance, there are different ways of providing forward guidance. It is difficult to evaluate the recent approach of publishing projected paths for the central bank's policy rate, given that we have gained only very limited experience to date. Another open issue relates to the transmission of central bank communication. The role of the media has barely been studied. Finally, while much of the empirical research has focused on the effects of communication on financial markets, a better understanding of the communication with the general public is required, since it is the general public whose inflation expectations eventually feed into the actual evolution of inflation – for example, through corresponding wage claims and savings, investment and consumption decisions.

See Also

- ▶ Bank of England
- ▶ European Central Bank
- ▶ Central Bank Independence
- ▶ Federal Reserve System
- ▶ Inflation
- ▶ Inflation Targeting
- ▶ Taylor Rules

Bibliography

Ball, L., and N. Sheridan. 2005. Does inflation targeting matter? In *The inflation-targeting debate*, ed. B. Bernanke and M. Woodford, 249–276. Chicago: University of Chicago Press.

Benati, L. 2008. Investigating inflation persistence across monetary regimes. *Quarterly Journal of Economics* 123: 1005–1060.

Bernanke, B. 2004. *Fedspeak*. Available at http://www.federalreserve.gov/boarddocs/speeches/2004/20040102/default.htm. Accessed 3 Jan 2004.

Bank for International Settlements (BIS). 2004. *74th annual report*. Basel: BIS.

Blinder, A. 2004. *The quiet revolution: Central banking goes modern*. New Haven: Yale University Press.

Blinder, A., M. Ehrmann, M. Fratzscher, J. de Haan, and D.-J. Jansen. 2008. Central bank communication and monetary policy: A survey of the evidence. *Journal of Economic Literature* 46(4): 910–945.

Dale, S., A. Orphanides, and P. Osterholm. 2008. *Imperfect central bank communication: Information versus distraction*, IMF working paper 08/60. Washington, DC: IMF.

Demiralp, S., and O. Jorda. 2002. The announcement effect: Evidence from open market desk data. *Federal Reserve Bank of New York Economic Policy Review* 8: 29–48.

Ehrmann, M., and M. Fratzscher. 2007. Communication by central bank committee members: Different strategies, same effectiveness? *Journal of Money, Credit, and Banking* 39: 509–541.

Ehrmann, M., and M. Fratzscher. 2009. Purdah: On the rationale for central bank silence around policy meetings. *Journal of Money, Credit, and Banking* 41(2–3): 517–527.

Gürkaynak, R., B. Sack, and E. Swanson. 2005. Do actions speak louder than words? The response of asset prices to monetary policy actions and statements. *International Journal of Central Banking* 1: 55–93.

Gürkaynak, R., A. Levin, and E. Swanson. 2009. Does inflation targeting anchor long-run inflation expectations? Evidence from long-term bond yields in the US, UK and Sweden. *Journal of the European Economic Association* (forthcoming).

Jansen, D.-J., and J. De Haan. 2006. Look who's talking: ECB communication during the first years of EMU. *International Journal of Finance and Economics* 11: 219–228.

Kahneman, D. 2003. Maps of bounded rationality: Psychology for behavioral economics. *American Economic Review* 93: 1449–1475.

King, M. 2005. *Monetary policy: Practice ahead of theory*. Available at http://www.bankofengland.co.uk/publications/speeches/2005/speech245.pdf. Accessed 17 May 2005.

Kohn, D., and B. Sack. 2004. Central bank talk: Does it matter and why? In *Macroeconomics, monetary policy, and financial stability*, ed. Bank of Canada, 175–206. Ottawa: Bank of Canada.

Kuttner, K., and A. Posen. 1999. *Does talk matter after all? Inflation targeting and Central Bank behavior*. Federal Reserve Bank of New York Staff Report 88. New York: Federal Reserve Bank.

Lange, J., B. Sack, and W. Whitesell. 2003. Anticipations of monetary policy in financial markets. *Journal of Money, Credit, and Banking* 35: 889–909.

Levin, A., F. Natalucci, and J. Piger. 2004. The macroeconomic effects of inflation targeting. *Federal Reserve Bank of St. Louis Review* 86: 51–80.

Morris, S., and H.S. Shin. 2002. Social value of public information. *American Economic Review* 92: 1521–1534.

Reeves, R., and M. Sawicki. 2007. Do financial markets react to Bank of England communication? *European Journal of Political Economy* 23: 207–227.

Svensson, L.E.O. 2003. What is wrong with Taylor rules? Using judgment in monetary policy through targeting rules. *Journal of Economic Literature* 41: 427–477.

Svensson, L.E.O. 2006. Social value of public information: Morris and Shin (2002) is actually pro transparency, not con. *American Economic Review* 96: 448–451.

Woodford, M. 2005. Central-Bank communication and policy effectiveness. In *The Greenspan era: Lessons for the future*, ed. Federal Reserve Bank of Kansas City, 399–474. Kansas City: The Federal Reserve Bank of Kansas City.

The views expressed in this article do not necessarily coincide with those of the European Central Bank or the Eurosystem.

Central Bank Independence

Carl E. Walsh

Abstract

Many countries have implemented reforms designed to grant their monetary authorities greater independence from direct political influence. These reforms were justified by research showing central bank independence was negatively correlated with average inflation among developed economies. An important line of research developed measures of central bank independence and studied their relationship with inflation and real economic activity. Different theoretical approaches have been used to model central bank independence. Critics of the reform movements towards central bank independence have expressed concerns that independence can weaken the accountability of central banks.

Keywords

Bank of England; Central bank independence; Central banks; European Central Bank (ECB); Federal Reserve System; Inflation; Inflation targeting; Monetary policy; Price stability; Rational expectations; Reserve Bank of New Zealand; Rogoff-conservative central bank

JEL Classifications

E52; E58

Central bank independence refers to the freedom of monetary policymakers from direct political or governmental influence in the conduct of policy.

During the 1970s and early 1980s, major industrialized economies experienced sustained periods of high inflation. To explain these periods of inflation, one must account for why central banks allowed them to happen. One influential line of argument pointed to the inflation bias inherent in discretionary monetary policy if the central bank's objective for real output (unemployment) is above (below) the economy's natural equilibrium level or if policymakers simply prefer higher output levels (Barro and Gordon 1983). Under rational expectations, the public anticipates that the central bank will attempt to expand the economy; as a consequence, real output is not systematically affected but average inflation is left inefficiently high.

This explanation for inflation raises the question why central banks might prefer economic expansions or have unrealistic output goals. Economists have frequently pointed to political pressures as the answer. Elected officials may be motivated by short-run electoral considerations, or may value short-run economic expansions highly while discounting the longer-run inflationary consequences of expansionary policies. If the ability of elected officials to distort monetary policy results in excessive inflation, then countries whose central banks are independent of such pressure should experience lower rates of inflation. Beginning with Bade and Parkin (1988), an important line of research focused on the relationship between the central bank and the elected government as a key determinant of inflation.

This empirical research found that average inflation was negatively related to measures of central bank independence. Cukierman (1992) provides an excellent summary of the empirical work; references to the more recent literature can be found in Eijffinger and de Haan (1996) and Walsh (2003, ch. 8). The empirical findings led to a significant body of work addressing the following questions: what do we mean by central bank independence? How should central bank independence be measured? What causal interpretation should be placed on the empirical correlations between central bank independence and macroeconomic outcomes discovered in the data? What is the theoretical explanation for these correlations?

The Meaning of Independence

The historical, legal and de facto relationships between a country's government and its central bank are very complex, involving many difference aspects. These include, but are not limited to, the role of the government in appointing (and dismissing) members of the central bank governing board, the voting power (if any) of the government on the board, the degree to which the central bank is subject to budgetary control by the government, the extent to which the central bank must lend to the government, and whether there are clearly defined policy goals established in the central bank's charter.

Most discussions have focused on two key dimensions of independence. The first dimension encompasses those institutional characteristics that insulate the central bank from political influence in defining its policy objectives. The second dimension encompasses those aspects that allow the central bank to freely implement policy in pursuit of monetary policy goals. Grilli et al. (1991) called these two dimensions 'political independence' and 'economic independence'. The more common terminology, however, is due to Debelle and Fischer (1994), who called these two aspects 'goal independence' and 'instrument independence'. Goal independence refers to the central bank's ability to determine the goals of

policy without the direct influence of the fiscal authority. In the United Kingdom, the Bank of England lacks goal independence since its inflation target is set by the government. In the United States, the Federal Reserve's goals are set in its legal charter, but these goals are described in vague terms (for example, maximum employment), leaving it to the Fed to translate these into operational goals. Thus, the Fed has a high level of goal independence. Price stability is mandated as the goal of the European Central Bank (ECB), but the ECB can choose how to interpret this goal in terms of a specific price index and definition of price stability.

Instrument independence refers only to the central bank's ability to freely adjust its policy tools in pursuit of the goals of monetary policy. The Bank of England, while lacking goal independence, has instrument independence; given the inflation goal mandated by the government, it is able to set its instruments without influence from the government. Similarly, the inflation target range for the Reserve Bank of New Zealand is set in its Policy Targets Agreement (PTA) with the government, but, given the PTA, the Reserve Bank has the authority to sets its instruments without interference. The Federal Reserve and the ECB have complete instrument independence.

Measuring Independence

The most widely employed index of central bank independence is due to Cukierman et al. (1992), although alternative measures were developed by Bade and Parkin (1988) and Alesina et al. (1991), among others.

The Cukierman, Webb and Neyapti index is based on four legal characteristics as described in a central bank's charter. First, a bank is viewed as more independent if the chief executive is appointed by the central bank board rather than by the government, is not subject to dismissal, and has a long term of office. These aspects help insulate the central bank from political pressures. Second, independence is greater the more policy decisions are made independently of government involvement. Third, a central bank is more independent if its charter states that price stability is the sole or primary goal of monetary policy. Fourth, independence is greater if there are limitations on the government's ability to borrow from the central bank.

Cukierman, Webb and Neyapti combine these four aspects into a single measure of legal independence. Based on data from the 1980s, they found Switzerland to have the highest degree of central bank independence at the time, closely followed by Germany. At the other end of the scale, the central banks of Poland and the former Yugoslavia were found to have the least independence.

Legal measures of central bank independence may not reflect the actual relationship between the central bank and the government. In countries where the rule of law is less strongly embedded in the political culture, there can be wide gaps between the formal, legal institutional arrangements and their practical impact. This is particularly likely to be the case in many developing economies. Thus, for developing economies, it is common to supplement or even replace measures of central bank independence based on legal definitions with measures that reflect the degree to which legally established independence is honoured in practice. Based on work by Cukierman, measures of actual central bank governor turnover, or turnover relative to the formally specified term length, are often used to measure independence. High actual turnover is interpreted as indicating political interference in the conduct of monetary policy.

Empirical Evidence

The 1990s saw many countries, both developed and developing, adopt reforms that increased central bank independence. This trend was strongly influenced by empirical analysis of the relationship between central bank independence and macroeconomic performance. Among developed economies, central bank independence was found to be negatively correlated with average inflation. The estimated effect of independence on inflation was statistically and economically

significant. Based on data from the high inflation years of the 1970s, for example, moving from the status of the Bank of England prior to the 1997 reforms that increased its independence to the level of independence then enjoyed by the Bundesbank would be associated with a drop in annual average inflation of four percentage points.

The form of independence may also matter for inflation. Debelle and Fischer (1994) report evidence that it is the combination of goal *dependence* and instrument *independence* that produces low average inflation, although their empirical results were weak.

Even if central bank independence leads to lower inflation, the case for independence would be greatly weakened if it also leads to greater real economic instability. However, little relationship was found between measures of real economic activity and central bank independence (Alesina and Summers 1993). In other words, countries with more independent central banks enjoyed lower average inflation rates yet suffered no cost in terms of more volatile real economic activity. Central bank independence appeared to be a free lunch.

While standard indices of central bank independence were negatively associated with inflation among developed economies, this was not the case among developing economies. Developing countries that experienced rapid turnover among their central bank heads tended to experience high rates of inflation. This is a case, however, in which causality is difficult to establish; is inflation high because of political interference that leads to rapid turnover of central bank officials? Or are central bank officials tossed out because they can't keep inflation down?

The empirical work attributing low inflation to central bank independence has been criticized along two dimensions. First, studies of central bank independence and inflation often failed to control adequately for other factors that might account for cross-country differences in inflation experiences. Countries with independent central banks may differ in ways that are systematically related to average inflation. After controlling for other potential determinants of inflation, Campillo and Miron (1997) found little additional role for central bank independence.

Second, treating a country's level of central bank independence as exogenous may be problematic. Posen (1993) has argued strongly that both low inflation and central bank independence reflect the presence of a strong constituency for low inflation. Average inflation and the degree of central bank independence are jointly determined by the strength of political constituencies opposed to inflation; in the absence of these constituencies, simply increasing a central bank's independence may not cause average inflation to fall.

Theoretical Models of Independence

Central bank independence has often been represented in theoretical models by the weight placed on inflation objectives. When the central bank's weight on inflation exceeds that of the elected government, the central bank is described as a Rogoff-conservative central bank (Rogoff 1985). This type of conservatism accorded with the notion that independent central banks are more concerned than the elected government with maintaining low and stable inflation. Rogoff's formulation reflects a form of both goal independence – the central bank's goals differ from those of the government – and instrument independence – the central bank is assumed to be free to set policy to achieve its own objectives. Because the central bank cares more about achieving its inflation goal, the marginal cost of inflation is higher for the central bank than it would be for the government. As a consequence, equilibrium inflation is lower.

One problem with interpreting independence in terms of Rogoff-conservatism is that Rogoff's model implies that a conservative central bank will allow output to be more volatile in order to keep inflation stable. Yet the empirical research finds no relationship between real fluctuations and measures of central bank independence.

An alternative way to model central bank independence is to view the central bank as having its own objectives, but the central bank must also take into account the government's objectives when deciding on policy. The central bank might

have either a lower desired inflation target than the government or an output target that, unlike the government's target, is consistent with the economy's natural rate of output. If actual policy is set to maximize a weighted average of the central bank's and the government's objectives, the relative weight on the central bank's own objectives provides a measure of central bank independence. With complete independence, no weight is placed on the government's objectives; with no independence, all weight is placed on the government's objectives. If the objectives of the central bank and the government differ only in their desired inflation target, then the degree of central bank independence affects average inflation but not the volatility of either output or inflation. Such a formulation is consistent with the empirical evidence discussed above.

Often, theoretical approaches have not distinguished clearly between goal and instrument independence. Suppose independence is measured by the relative weight on the government's and the central bank's objectives. This can be interpreted as reflecting either goal dependence – the objectives of the central bank must put some weight on the goals of the government – or instrument dependence – the actual instrument setting diverges from what would be optimal from the central bank's perspective in order to reflect the government's concerns.

Independence and Accountability

While many countries have granted their central banks more independence, the idea that central banks should be completely independent has come under criticism. This criticism focuses on the danger that a central bank that is independent will not be accountable. Although maintaining low and stable inflation is an important societal goal, it is not the only macroeconomic goal; monetary policy may have no long-run effect on real economic variables, but it can affect the real economy in the short run. In a democracy, delegating policy to an independent agency requires some mechanism to ensure accountability. For this reason, reforms have often granted central banks instrument independence while preserving a role for the elected government in establishing the goals of policy and in monitoring the central bank's performance in achieving these goals.

See Also

- ▶ Inflation
- ▶ Inflation Targeting
- ▶ Optimal Fiscal and Monetary Policy (Without Commitment)

Bibliography

Alesina, A., and L. Summers. 1993. Central bank independence and macroeconomic performance. *Journal of Money, Credit, and Banking* 25: 157–162.

Bade, R., and M. Parkin. 1988. *Central bank laws and monetary policy*. Working paper. Department of Economics, University of Western Ontario.

Barro, R., and D. Gordon. 1983. A positive theory of monetary policy in a natural-rate model. *Journal of Political Economy* 91: 589–610.

Campillo, M., and J. Miron. 1997. Why does inflation differ across countries? In *Reducing inflation: Motivation and strategy*, ed. C. Romer and D. Romer. Chicago: University of Chicago Press.

Cukierman, A. 1992. *Central bank strategy, credibility, and independence: Theory and evidence*. Cambridge, MA: MIT Press.

Cukierman, A., S. Webb, and B. Neyapti. 1992. Measuring the independence of central banks and its effects on policy outcomes. *World Bank Economic Review* 6: 353–398.

Debelle, G., and S. Fischer. 1994. How independent should a central bank be? In *Goals, guidelines and constraints facing monetary policymakers*, ed. J. Fuhrer. Boston: Federal Reserve Bank of Boston.

Eijffinger, S., and J. de Haan. 1996. *The political economy of central-bank independence*. Special Papers in International Economics, No. 19. Princeton University.

Grilli, V., D. Masciandaro, and G. Tabellini. 1991. Political and monetary institutions and public financial policies in the industrial countries. *Economic Policy* 6: 341–392.

Posen, A. 1993. Why central bank independence does not cause low inflation: There is no institutional fix for politics. In *Finance and the international economy*, vol. 7, ed. R. O'Brien. Oxford: Oxford University Press.

Rogoff, K. 1985. The optimal commitment to an intermediate monetary target. *Quarterly Journal of Economics* 100: 1169–1189.

Walsh, C. 2003. *Monetary theory and policy*, 2nd ed. Cambridge, MA: MIT Press.

Central Banking

Charles Goodhart

When the first government-sponsored banks were founded in Europe, for example the Swedish Riksbank (1668) and the Bank of England (1984), there was no intention that these should undertake the functions of a modern central bank, that is, discretionary monetary management and the regulation and support, for example through the 'lender of last resort' function, of the banking system. Instead, the initial impetus was much more basic, generally relating to the financial advantages a government felt that it could obtain from the support of such a bank, whether a State bank, as in the case of the Prussian State Bank, or a private bank, like the Bank of England. This naturally involved some favouritism, often supported by legislation, by the government for this particular bank in return for its financial assistance. The favoured bank was often granted a monopoly advantage, for example over the note issue in certain areas, or as the sole chartered joint stock bank in the country; and this may have had the effect in some countries, such as England and France, of weakening the early development of other commercial banks, so that, at the outset, the foundation of a government-sponsored bank was a mixed blessing for the development of banking in such countries.

Other government-sponsored central banks, for example the Austrian National Bank founded in 1816 at the end of the Napoleonic wars, were established to restore the value of the national currency, notably after its value had been wrecked by government over-issue in the course of war finance. Others were founded partly in order to unify what had become in some cases (e.g in Germany, Switzerland and Italy) a somewhat chaotic system of note issue; to centralize, manage and protect the metallic reserve of the country, and to facilitate and improve the payments system. While these latter functions were seen as having beneficial economic consequences, the ability to share in the profits of seignorage and greater centralized control over the metallic (gold) reserve had obvious political attractions as well. In any case, prior to 1900, most economic analysis of the role of Central Banks concentrated on the question of whether the note issue, and the gold reserves of the country, should be centralized, and, if and when centralized, how controlled by the Central Bank.

Once such government-sponsored banks had been established, however, their central position within the system, their 'political' power as the government's bank, their command (usually) over the bulk of the nation's specie reserve, and, most important, their ability to provide extra cash, notes, by rediscounting commercial bills made them become the bankers' bank: commercial banks would not only hold a large proportion of their own (cash) reserves as balances with the Central Bank, but also rely on it to provide extra liquidity when in difficulties. In several early cases, such as the Bank of England's, this latter role had not been initially intended; in most cases of Central Banks founded in the 19th century the full ramifications of their role as bankers' bank were only dimly perceived at the time of their founding; these functions developed naturally from the context of relationships within the system.

Initially, indeed, the role of Central Banks in maintaining the convertibility of their notes, into gold or silver, was not different, nor seen as different, from that of any other bank. Their privileged legal position, as banker to the government and in note issue, then led naturally to a degree of centralization of reserves within the banking system in the hands of the Central Bank, so it became a banker's bank. It was the responsibility that this position was found to entail, in the process of historical experience, that led Central Banks to develop their particular art of discretionary monetary management and overall support and responsibility for the health of the banking system at large.

This management has had two (interrelated) aspects: a macro function and responsibility relating to overall monetary conditions in the economy, and a micro function relating to the health and wellbeing of the (individual) members of the

banking system. Until 1914 such management largely consisted of seeking to reconcile the need to maintain the chosen metallic standard, usually the gold standard, on the one hand with concern for the stability and health of the financial system, and beyond that of the economy more widely, on the other. Thereafter, as the various pressures of the 20th century disrupted first the gold standard and thereafter the Bretton Woods' system of pegged exchange rates, the macroeconomic objectives of monetary management have altered and evolved. Yet at all times concern for the health of the banking system has remained a paramount concern for the Central Bank.

This concern for the wellbeing of the banking system as a whole was, at least for those Central Banks founded in the 19th century or before, largely an evolutionary development and not one that they had been programmed to undertake from the start. Indeed in England the legislative framework of the 1844 Bank Charter Act was to prove something of a barrier to the development of the micro-supervisory functions of the Bank: for this Act divided the Bank into two Departments – the Issue Department, whose note issuing function was to be closely constrained by strict rules (to maintain the Gold Standard); and the Banking Department, which was intended to behave simply as an ordinary competitive, profit-maximizing, commercial bank.

Nevertheless the micro-functions of a Central Bank in providing a central (and therefore economical) source of reserves and liquidity to other banks, and hence both a degree of insurance and supervision, cannot be undertaken effectively by a commercial competitor, basically because of competitive conflicts of interest. The advantages of having some institutions providing such micro-Central Banking functions are such that even in those various countries initially without Central Banks there was some natural tendency towards their being provided, after a fashion, from within the private sector – for example by clearing houses in the United States, or by a large commercial bank providing quasi-Central Bank functions. Nevertheless, because of conflicts of interest, such functions were not, and cannot be, adequately provided by competing commercial institutions.

Some Central Banks, mainly those that began their existence under private ownership (e.g. the Bank of England, the Banca d'ltalia, but also some that were subject to political oversight, e.g. the Banque de France, the Commonwealth Bank of Australia), retained for a considerable time a large role in ordinary commercial banking. It was, however, the metamorphosis from their involvement in commercial banking, as a competitive, profit-maximizing bank among many, to a non-competitive, non-profit-maximizing role that marked the true emergence in those countries of proper Central Banking. This metamorphosis occurred naturally, but with considerable difficulty in England, the difficulty arising in part from the existence of property rights in the profits of the Bank, and in part from concern about the moral hazards of the Bank consciously adopting a supervisory role, (as evidenced in the arguments between Bagehot and Hankey, reported in Bagehot's *Lombard Street*).

Indeed, with the Central Bank coming to represent the ultimate source of liquidity and support to the individual commercial banks, this micro-function does bring with it naturally a degree of 'insurance'. Such insurance, in turn, does involve some risk of moral hazard: commercial banks, believing that they will be protected by their Central Bank from the consequences of their own follies, may adopt too risky and careless strategies. That concern has led Central Banks to become involved – to varying extents – in the regulation and supervision of their banking systems. In all countries the Central Bank plays *some* role in the support of its commercial banks, because it alone can provide 'lender of last resort' assistance; but the extent to which it shares the insurance, supervisory, and regulatory function, both for the banking system more narrowly and for the wider financial system, with government and private bodies set up specifically for such purposes, varies from country to country. With structural changes apparently breaking down the barriers between the banking system on the one hand and other financial intermediaries on the other in the course of the 1970s and 1980s, the question of the division of responsibility of the Central Bank on the one hand, and other

supervisory government bodies and insurance agencies on the other, has become topical.

The Central Bank's more glamorous function is the conduct of macro-monetary policy. The main objective of this function in normal times has been to maintain the (internal and external) value, and reputation, of the national currency. At times of national crisis, notably during wars, however, the financial needs of the State have generally overridden the desire for financial stability, with the conduct of monetary policy then being mainly determined by questions of how the necessary finance can most effectively be mobilized to support the urgent needs of the State. Apart from such national emergencies, the desire to achieve financial stability became synonymous, during the 19th and early 20th centuries, with adherence to the Gold Standard.

The break-down of the Gold Standard in the interwar period left many countries with high unemployment, a falling price level, and international trade and capital flows increasingly constrained by direct controls. In this context it became widely felt that monetary policy was relatively powerless: once interest rates were brought down to low levels, there was little more, it was argued, that monetary policy could do. The management of aggregate demand would, therefore, have to be left to fiscal policy, with direct controls of various kinds used to constrain subsequent inflationary pressures (e.g. in World War II) and international disequilibria.

The erosion of direct controls in the late 1940s and 1950s, and the establishment of the Bretton Woods system of pegged, but adjustable, exchange rates, meant that Central Banks generally were able, during the 1950s and 1960s, to return to their accustomed policy of maintaining the value of their national currencies by seeking to hold these pegged to the US dollar and thence, until the late 1960s, to gold. With the US dollar at the centre of the world financial system, the Federal Reserve System had a different and special responsibility, to maintain the internal stability of the $. After many successful years, US monetary policy and the Bretton Woods system were overwhelmed by pressures arising from the Vietnam War, political strains within the Western Alliance, and, finally, the 1973 Oil Shock.

Up till then, most Western governments had sought to maximize employment and growth, along broadly Keynesian lines, subject to trying to maintain the exchange rate peg. With that peg no longer in place after 1972, governments then placed various emphases on supporting full employment on the one hand and monetary constraint on the other. In the event, however, there seemed no evidence that countries with more expansionary monetary policies, and thence more inflation, did achieve notably higher rates of growth of employment. This experience led directly to the adoption of 'pragmatic' monetarist policies by the Central Banks of the main industrialized countries, whereby they sought to achieve publicly announced, steadily declining rates of growth for certain domestic monetary intermediate target aggregates.

This policy shift has, in turn, had a chequered history. Monetarists claim that the commitment to, and technical execution of, monetary targetting has been unsatisfactory. Keynesians claim that it has involved no more than simple deflation, with the policy's success in reducing inflation in the early 1980s tarnished by a dramatic growth in unemployment and a poor rate of growth of real output. Moreover, the conduct of policy has been complicated by a generally growing instability, partly induced by structural change, in the relationship between money and nominal incomes, an unstable velocity of money; and also by serious and persistent volatility in exchange rates and interest rates, often leaving these seemingly way out-of-line with economic fundamentals.

As of 1985, it seems difficult to see how a fully international system of pegged exchange rates could be re-established, though this would provide the traditional, and simplest, milieu for Central Bank policy. (This, though, would still allow regional groupings of countries to seek to maintain a stable exchange rate system between themselves, such as the European Monetary System, generally based on a central key currency within the group.) On the other hand, previous enthusiasm for rules, and for fixed targets for monetary growth, is dissipating, partly as the evolving

structure of the financial system once again brings into question the appropriate definition, role, and essential properties, of money and banks. So for the moment, there seems no valid alternative to a discretionary conduct of monetary policy, with an eye not only both to monetary and exchange rate developments, but also to the broader evolution of the economy.

See Also

- ▶ Bank Rate
- ▶ Cheap Money
- ▶ Dear Money
- ▶ Financial Intermediaries
- ▶ Monetary Policy

References

Classical

Bagehot, W. 1873. *Lombard street*. London: Henry S. King.

Fetter, F. 1965. *Development of British monetary orthodoxy, 1797–1875*. Cambridge, MA: Harvard University Press.

Thornton, H. 1802. *An inquiry into the nature and effects of the paper credit of Great Britain*. London: Hatchard.

Evolution

Goodhart, C.A.E. 1985. *The evolution of central banks*, ICERD monograph. London: London School of Economics.

Hawtrey, R.G. 1932. *The art of central banking*. London: Longmans.

Sayers, R. 1957. *Central banking after Bagehot*. Oxford: Clarendon Press.

Smith, V. 1936. *The rationale of central banking*. London: P.S. King & Son.

Timberlake Jr., R. 1978. *The origins of central banking in the United States*. Cambridge, MA: Harvard University Press.

US National Monetary Commission. 1910–11. *Twenty volumes of papers and original material on banking and central banking in all major industrialized countries*. Washington, DC: Government Printing Office.

Veit, O. 1969. *Grundriss der Wahrungspolitik*, 3rd ed. Frankfurt: Fritz Knapp Verlag.

Contemporary

Bank of England. 1984. *The development and operation of monetary policy, 1960–1983*. Oxford: Clarendon Press.

Board of Governors of the Federal Reserve System. *The federal reserve system: Purposes and functions*. Washington, DC: Federal Reserve Board.

Duwendag, D., et al. 1985. *Geldtheorie und Geldpolitik*, 3rd ed. Cologne: Bund-Verlag.

Federal Reserve Bank of New York. 1983. *Central bank views on monetary targeting*. New York: Federal Reserve Bank of New York.

Meek, P. 1982. *US monetary policy and financial markets*. New York: Federal Reserve Bank of New York.

Woolley, J. 1984. *Monetary politics*. Cambridge: Cambridge University Press.

Central Limit Theorems

Werner Ploberger

Abstract

Central limit theorems describe the behaviour of distributions of sums of random variables. We start with the classical result of distributions of sums of independent random variables converging to the Gaussian (bell-curve) distribution. We describe the most important cases of convergence to Gaussian distributions (sums of martingale differences) as well as convergence to other distributions.

Keywords

Central limit theorems; Convergence; Edgeworth expansions; Feller condition; Laplace, P. S.; Lindeberg condition; Long-term variance; Lyapunov condition; Martingale differences; Maximum likelihood; Monte Carlo simulation

JEL Classifications

C10

At the end of the 17th century, the mathematician Abraham de Moivre first used the normal distribution as an approximation for the percentage of successes in a large number of experiments. Later on, Laplace generalized his results, but it took 20th century mathematics to give an exact and

complete description of this subject. So let me now describe the modern approach. We assume that for each n we have given a sequence $X_{1,n,}...,$ $X_{n,n}$ of random variables, which we assume to be independent. Then we want to 'approximate' the distribution of

$$S_n = \sum_{i=1}^{n} X_{i,n}$$

by a standard normal distribution, whose density equals

$$\frac{1}{\sqrt{2\pi}} \int_A \exp\left(-\frac{x^2}{2}\right) \mathrm{d}x.$$

Let us denote by $P(B)$ the probability of an event B. If X is a random variable, than let us denote by $E(X)$ its expectation. For $A \subseteq \mathrm{R}$ let $[X \in A]$ be the event that X takes a value into A. Written in formal terms, we want to establish that

$$\lim_{n\to\infty} P([S_n \in A]) = \frac{1}{\sqrt{2\pi}} \int_A \exp\left(-\frac{x^2}{2}\right) \mathrm{d}x \quad (1)$$

or

$$\lim_{n\to\infty} Ef(S_n) = \frac{1}{\sqrt{2\pi}} \int f(x) \exp\left(-\frac{x^2}{2}\right) \mathrm{d}x. \quad (2)$$

The first question we have to ask ourselves is the nature of the approximation. Clearly it is impossible to approximate the distribution of S_n for *all* sets. Consider the binomial distribution discussed above. In this case, each S_n can only take a finite number of values. Therefore the possible values for all S_n lie for all n in a countable set, which has zero probability under the normal distribution.

So we have to aim at a compromise: the smaller the class of sets A or functions f, the more 'convergent' sequences S_n we have. The most successful compromise is the convergence in distribution of the random variables (or the weak convergence of the probability distributions). We postulate that (2) holds for all bounded, continuous functions f. This requirement can be shown to be equivalent to postulating that (1) holds for all sets A so that the boundary of A (that is, the difference between closure of A and inner points of A) has zero probability under the limiting measure. So in our case, where the limiting distribution is normal, (1) holds if A is an interval (a, b): the boundary consists of two points, namely a and b. Equation (1) does not hold if, for example, A is the set of all rational numbers in (0, 1): then the boundary equals [0, 1], which obviously has non-zero probability under the normal distribution (see Billingsley 1999).

It is noteworthy that there are many more equivalent ways to define convergence in distribution for unidimensional random variables; for example, convergence in distribution is equivalent to the convergence of the cumulative distribution functions to the cumulative distribution function of the limiting distribution in all points where the latter is continuous. Another well-known criterion is the convergence of the characteristic functions.

Now we are in a position to formulate our first main theorem, the central limit theorem (CLT) of Lindeberg and Feller (see Billingsley 1995).

Suppose we have given a triangle array of random variables $X_{i,n}$, so that for each n the $X_{i,n}$ are independent, not necessarily identically distributed. We furthermore have

$$EX_{i,n} = 0,$$
$$\sum_{i=1}^{n} \mathrm{Var}(X_{i,n}) = 1.$$

Then the following two propositions are equivalent:

- The 'Lindeberg' condition: For all $\delta > 0$

$$\sum_{i=1}^{n} E\left(X_{i,n}^2 I\left[|X_{i,n}| > \delta\right]\right) \quad (\mathrm{L})$$

converges to zero.

- Our sums

$$S_n = \sum_{i=1}^{n} X_{i,n}$$

converge in distribution to a standard normal and the 'Feller' condition is satisfied:

$$\max_{1\le i\le n} \mathrm{Var}(X_{i,n}) \to 0. \qquad \text{(F)}$$

It seems plausible to assume the Feller condition (F). It simply states that the maximal contribution of an individual $X_{i,n}$ to the variance of the sum gets arbitrarily small. This seems reasonable. The Lindeberg condition (L) which is necessary for our theorem is a little stronger. Not only the maximum, but the *total* contribution of the $X_{i,n}$ taking 'large' values to the variance of the sum, must vanish asymptotically!

It is quite easy to establish that (L) is fulfilled if

$$X_{i,n} = \frac{1}{\sqrt{1}} X_i, \qquad (3)$$

where the X_i are independent and identically distributed. In the general case, a sufficient condition is the 'Lyapunov condition': for some fixed $\varepsilon > 0$ we have

$$\sum_{i=1}^{n} E\left(|X_{i,n}|^{2+\varepsilon}\right) \to 0.$$

So we need a little more than second moments to establish convergence to a standard normal. Practitioners often assume that the requirements of the theorems are fulfilled automatically. This assumption is quite dangerous. We need a little more than lack of outliers; the contribution to the variance of the largest values must be negligible.

This relation between higher moments and goodness of the approximation with a standard normal is extensive. Under the assumption of at least three absolute moments, the theorem of Berry–Esseen shows that in the case (3) of independent, identically distributed X_i the maximal difference between the cumulative distribution functions of S_n and the standard normal is $1/\sqrt{n}$. Related are 'coupling' results. One can show that – possibly on a richer probability space – there exist *exactly* normally distributed random variables U_n. In particular, if the X_i have a Laplace transform, then the 'Hungarian construction' allows one to construct U_n so that the difference to S_n is $\mathrm{O}(\log(n)/\sqrt{n})$. If the X_i 'only' have fourth moments, then it is easy (for the insider: use Skorohod embedding) to construct U_n so that the difference to S_n is of the order of $1/\sqrt[4]{n}$.

All these bounds are very interesting from the theoretical point of view. Playing around with numbers for n with realistic sample sizes, one can easily see that the bounds found that way are unrealistic. Although these bounds cannot be improved, they are a little pessimistic. Nevertheless, they indicate when we venture into dangerous territory: a lack of fourth moments indicates a 'slow' convergence.

So the normal approximation is a useful first-order approximation of the distributions of sums of random variables. To improve this approximation, various techniques are used. Since the 19th century, Edgeworth expansions have proved useful. Nowadays, however, cheap computing makes direct calculation of distributions by Monte Carlo simulation possible.

Independent, Non-normal Limit Theorems

Let us define $\mathrm{X}_{i,n}$ to be independent, identically distributed and taking the value of zero with probability $1 - \lambda/\mathrm{n}$ and one with probability λ/n with some $\lambda > 0$. Now one has an easy example where the Lindeberg condition is not fulfilled. (For $\delta < 1$, $\sum_{i=n}^{n} E\left(X_{1,n}^2 I(|X_{i,1}| > \delta\right) = \lambda$, since $\mathrm{X}_{i,n}$ can take only the values 0 and 1). Nevertheless, it is well known that $\sum_{i=1}^{n} X_{i,n}$ converges in distribution to a Poisson distribution with intensity λ. So the normal distribution is not the only limiting distribution of sums of independent random variables. One can, however, show that the normal and the Poisson distribution and mixtures (with possibly an infinite number of components) of these distributions are the only possible limits of sums S_n of independent, identically distributed random variables $X_{i,n}$. These limiting distributions are called 'infinitely divisible'. A precise formula for the logarithm of the characteristic function is given by the formula of Levy–Khinchin.

We even have some analogon, some generalization of the normal distribution. The properly normalized sum of normally distributed random variables is normal again. Can we generalize this property? Let us assume that

$$X_{i,n} = a_n(X_i - b_n), \tag{4}$$

where the X_i are independent and identically distributed, and the a_n are scale factors, and let us assume that the distribution of the S_n is identical to the distribution of the X_i. These distributions are called the 'stable' distributions. Their density is determined essentially by two parameters, traditionally called α and β. α determines the 'tail behaviour' and varies between 0 and 2, and β determines the symmetry. For $\alpha = 2$, we have the normal distribution, for $\alpha < 2$ the distributions are more heavily tailed: in general, one has only moments of order smaller than α. There is no closed form for their densities in the general case, only the characteristic functions can be expressed by elementary functions. One special case ($\alpha = 1$) is the Cauchy distribution with density

$$\frac{1}{\pi(1+x^2)}.$$

The index α determines the scale factors a_n: in general, one has $a_n = n^{\frac{1}{\alpha}}$.

Convergence of sums to stable distributions can be achieved in more general circumstances. In general, under certain conditions on the 'tail' of the X_i (the probabilities exceeding 'large' values have to obey certain regularity conditions) the sums of the $X_{i,n}$ defined by (4) one can ensure convergence (see Ibragimov and Linnik 1971).

Central Limit Theorems for Dependent Random Variables

Many econometric applications involve sums of dependent random variables. Hence it is important to remove the requirement of independence. Traditionally, one tried to replace independence by some form of 'mixing'.

Independence of two σ-algebras $\mathfrak{A}$ and $\mathfrak{B}$ can be defined in various ways.

Usually one defines $\mathfrak{A}$ and $\mathfrak{B}$ to be independent if for all $A \in \mathfrak{A}$ and $B \in \mathfrak{B}$

$$P(A \cap B) = P(A)P(B).$$

Another usual definition is that for all $A \in \mathfrak{A}$

$$P(A/\mathfrak{B}) = P(A),$$

where $P(\cdot/\cdot)$ should denote the conditional probability. Consequently, one can measure the 'degree of dependence' of σ-algebras $\mathfrak{A}$ and $\mathfrak{B}$ by

$$\alpha(\mathfrak{A}, \mathfrak{B}) = \sup_{A \in \mathfrak{A}, B \in \mathfrak{B}} |P(A \cap B) - P(A)P(B)|$$

or

$$\psi(\mathfrak{A}, \mathfrak{B}) = \sup_{A \in \mathfrak{A}} |P(A/\mathfrak{B}) - P(A)|.$$

Suppose one has give a process X_t. Then one defines the 'mixing coefficients'

$$\alpha_k = \sup_t \alpha\left(\mathfrak{A}_\sigma\left(X_t, X_{t+1,\ldots}\right), \mathfrak{A}_\sigma\left(X_{t-k}, X_{t-1-k,\ldots}\right)\right)$$

or

$$\psi_k = \sup_t \psi\left(\mathfrak{A}_\sigma\left(X_t, X_{t+1,\ldots}\right), \mathfrak{A}_\sigma\left(X_{t-k}, X_{t-1-k,\ldots}\right)\right).$$

Typically, conditions Like

$$\sum \sqrt{\alpha_k} < \infty$$

or

$$\psi_k \to 0$$

are sufficient conditions for a CLT. So the CLT remains valid for stationary processes if the random variables in questions get less and less dependent if the time difference gets larger and larger (Ibragimov and Linnik 1971; Davidson 1994).

CLT for Martingale Differences

One of the most important applications is the CLT for martingale differences. A process X_t is a 'martingale difference' if for all t

$$E(X_t/\mathfrak{F}_{t-1}) = 0,$$

where $\mathfrak{F}_{t-1}$ is an increasing sequence of σ-algebras which contain at least $X_{t-1}, X_{t-2},\ldots$. Then we have a result perfectly analogous to the case of independent random variables.

Suppose we have given a triangle array $X_{t,T}, t = 1,\ldots, T$, of martingale differences with σ-algebras $\mathfrak{F}_{t-1,T}$ and the following two conditions are satisfied:

- the conditional Lindeberg condition

$$\sum_{t=1}^{T} E\left(X_{t,T}^2 I_{\left[|X_{t,T}|\geq\varepsilon\right]}/\mathfrak{F}_{t-1,T}\right) \to 0,$$

- the norming condition

$$\sum_{t=1}^{T} E\left(X_{t,T}^2/\mathfrak{F}_{t-1,T}\right) \to 1,$$

where the convergence should be understood to be in probability. Then

$$S_n = \sum_{i=1}^{n} X_{i,n}$$

converges in distribution to a standard normal distribution (Davidson 1994; Hall and Heyde 1980).

This limit theorem is one of the most important ones for applications in econometrics. It is relatively easily seen that derivatives of log-likelihood functions are martingale differences. Hence this theorem is instrumental in establishing the limit theorems for maximum likelihood estimators.

An easy consequence of the theorem is that for every (strictly) stationary, ergodic martingale difference X with $\sigma^2 = E\left(X_t^2\right) < \infty$ we have an almost classical CLT:

$$\frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} X_i$$

which converges in distribution to a standard normal.

Gordin's Theorem

Martingale differences form a large class of processes. Unfortunately, however, this class is not sufficiently large for many important applications (martingale differences must be, for example, uncorrelated). As an alternative, one might use mixing conditions. These conditions are, however, hard to verify. They usually involve inequalities involving *all* events from the σ-algebras involved. Hence a theorem allowing for general, autocorrelated processes with conditions which are easy to verify is an important tool in theoretical econometrics. Such a result was found by Gordin in 1969. Hayashi (2000) demonstrates the versatility of the theorem.

Suppose we have a stationary, ergodic process X_i, $i \in Z$ so that $EX_i^2 < \infty$. Assume that $\mathfrak{F}_i$ are adapted σ-algebras (that is, X_i are $\mathfrak{F}_i$-measurable), and let

$$\varepsilon_i = E(X_i/\mathfrak{F}_1) - E(X_i/\mathfrak{F}_0).$$

Then let us assume that

$$\sum_{i=1}^{\infty} \sqrt{E\varepsilon_i^2} < \infty.$$

Then

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i$$

converges in distribution to a normal distribution with zero mean and variance σ_{LT}^2 where

$$\sigma_{LT}^2 = E\left(\sum_{i=1}^{\infty} \varepsilon_i\right)^{2'}.$$

σ_{LT}^2 is usually called the 'long-term variance'.

Conclusion

Almost all theorems about limit distributions of estimators and test statistics depend on central limit theorems. So it should not be surprising that central limit theorems and their generalizations are an active field of research. Especially, generalizations of the concept of convergence in distribution to more general spaces generate theorems, which are important from the theoretical as well as the practical point of view. Billingsley (1999) and Davidson (1994) give an introduction to these 'functional limit theorems'.

See Also

► Functional Central Limit Theorems

Bibliography

Billingsley, P. 1995. *Probability and measure*. 3rd ed. New York: Wiley.
Billingsley, P. 1999. *Convergence of probability measures*. 2nd ed. New York: Wiley-Interscience.
Davidson, J. 1994. *Stochastic limit theory: An introduction for econometricians*. Oxford: Oxford University Press.
Hall, P., and C.C. Heyde. 1980. *Martingale limit theory and its application*. New York: Academic.
Hayashi, F. 2000. *Econometrics*. Princeton: Princeton University Press.
Ibragimov, I.A., and Yu.V. Linnik. 1971. *Independent and stationary sequences of random variables*. Groningen: Wolters-Noordhoff.

Central Place Theory

Marcus Berliant

Abstract

Central place theory is a descriptive theory of market area in a spatial context. Its main assumptions are that consumer population is distributed uniformly while firms locate in cities; the latter form a hierarchy with overlapping market areas. But central place theory runs afoul of Starrett's spatial impossibility theorem. Not grounded in the analytical tools of modern economics, central place theory does not have firm foundations. Thus, it is difficult to build on central place theory, either theoretically or empirically.

Keywords

Central place theory; City hierarchy; Increasing returns to scale; Krugman, P.; Spatial impossibility theorem; Urban agglomeration

JEL Classifications

R14

Central place theory is a descriptive theory of market area in a spatial context. Its definition, history, and relation to modern microeconomic theory are set out in this article.

Central place theory is a collection of loosely related, informal, descriptive models of city size, city location, and market area based on the trade-off between increasing returns to scale in production and the cost of transport of goods from firm to home. Land markets are often absent. At its core, central place theory is an empirically motivated description of production in southern Germany. It is a remarkable empirical regularity in search of a formal theory; a better name would be 'central place regularity'.

The beginnings of the theory are attributed to Christaller (1933), who first made detailed observations of urban hierarchies and then attempted to model them. The basic ideas put forward are that consumer population is distributed uniformly, while firms locate in cities. Cities form a hierarchy in that cities higher in the hierarchy produce all the goods that cities one level lower in the hierarchy produce, and one more. The ratio of market areas of a commodity produced only at a given level of the hierarchy (and above) to the market area of a commodity produced at the next lower level of the hierarchy (and above) is assumed to be constant, independent of the level in the hierarchy considered. Thus, the cities in a given area form a hierarchy where the size of a city's market area and the variety of commodities it offers are perfectly

correlated. In graphical terms, the result is a collection of hierarchically ordered cities with the market areas of cities not at the same level of the hierarchy overlapping, but market areas of cities at the same level disjoint. Commodities characterized by low transport cost but high returns to scale are provided by a few cities high in the hierarchy. Commodities characterized by high transport cost but low returns to scale are provided by most cities.

Lösch (1944) expanded on this theory. He postulated a homogeneous agricultural plane with farmers. Some turn to beer production, and face linear, downward sloping demand curves with choke prices, that is, prices above which the demand is for beer is zero. For a given price at the brewery, total delivered price increases with distance from the plant due to transport cost. In the plane with a uniform distribution of inebriated consumers or farmers, demand for a firm's beer is given by the volume of a cone centred at the brewery, with height given by the brewery's mill price and the slope of its sides determined by the demand curve and the cost of beer transport. With a marginal cost curve, equilibrium can be found. Unfortunately, the collection of bases of cones, namely, disks, does not partition the plane. So hexagons are used, forming a Teutonic triangulation of hierarchical hexagons. In this theory, the central places are the breweries. (St. Louis is a prime example).

One can view the theory as producing a complex of overlapping, ordered layers of hexagonal partitions of the plane corresponding to the market areas of cities in a hierarchy. Agriculture is the basis for and genesis of this structure.

The theory has developed beyond these basic descriptive models; see McCann (2001, ch. 2.7) for a nice summary and cites. Hartwick (2004) is the culmination of a line of research more in accord with optimizing behaviour, pricing, and trade theory that also relates the models to the rank-size rule.

The reader should be cautious in interpreting this entire literature because equilibrium and efficiency are often confused, while the models tend to be mechanistic in nature as opposed to allowing agents to optimize in equilibrium. To the general economist, the theory will appear to be informal and imprecise. Paul Krugman (1995, pp. 38–41) criticizes central place theory, or 'Germanic geometry', for its lack of formal foundations, particularly regarding market structure and firm behaviour. This criticism applies even to the contemporary literature. (Paul Krugman is also credited with the first alliteration in this literature. This article only builds on the original contribution).

Even if one is willing to overlook these defects, there is one further important flaw. Central place theory generally runs afoul of Starrett's spatial impossibility theorem; see Starrett (1978), Fujita (1986), and Fujita and Thisse (2002, ch. 2.3) for discussion. In essence, the impossibility theorem says that, in a closed economy with perfect and complete markets at all locations, location-independent utility and production functions, and no relocation cost, there is no competitive equilibrium where commodities are transported. Thus, if the assumptions are satisfied, either there is no equilibrium or in equilibrium agents and commodities are distributed uniformly among inhabited locations, and locations are autarkic. Central place theory apparently makes these assumptions, though due to its imprecision perhaps it doesn't. Naturally, although the literature considers consumer migration at times, the assumption of a uniform distribution of consumers could render the theorem inapplicable. *I* conjecture that it simply makes the existence of an (autarkic) equilibrium more likely. But this is probably not worth pursuing, as location models that fix consumer locations in a uniform distribution can generate only cities without people.

So where does this leave us? The modern theory of agglomeration, and thus the modern theory of central places, begins with the impossibility theorem. Its contrapositive tells us that, to generate models with non-trivial agglomeration at equilibrium, at least one of the hypotheses must be violated. Even then, equilibrium might not exist, or in equilibrium cities could collapse to a point or have agents spread uniformly. Models of non-trivial cities involve a very delicate balancing act between forces pulling agents together and forces

pushing them apart. The New Economic Geography has provided one of several possible types of models capable of producing cities and even hierarchies of cities. Fujita and Mori (1997) and Fujita et al. (1999) generate a form of central place theory in a general equilibrium framework by employing imperfect competition and increasing returns at the firm level. Unfortunately, this type of model has many defects, as detailed in Berliant (2006), including a reliance on specific functional forms and indeterminacy: one equilibrium is selected from a continuum.

Central place theory is not grounded in the analytical tools of modern economics, so it does not have firm foundations. Thus, it is difficult to build on central place theory, either theoretically or empirically.

In my view, the future of central place theory is as a stylized fact to be explained by our models, much like the rank-size rule.

See Also

- ▶ Spatial Economics
- ▶ Systems of Cities
- ▶ Urban Agglomeration
- ▶ Urban Economics

References

Berliant, M. 2006. Well isn't that spatial?! Handbook of regional and urban economics, vol. 4: A view from economic theory. *Journal of Economic Geography* 6: 107–110.

Christaller, W. 1933. *Central places in Southern Germany.* Englewood Cliffs: Prentice-Hall. 1966.

Fujita, M. 1986. Urban land use theory. In *Location theory*, ed. J. Lesourne and H. Sonnenschein. New York: Harwood Academic Publishers.

Fujita, M., and T. Mori. 1997. Structural stability and evolution of urban systems. *Regional Science and Urban Economics* 27: 399–442.

Fujita, M., and J.-F. Thisse. 2002. *Economics of agglomeration*. Cambridge: Cambridge University Press.

Fujita, M., P. Krugman, and T. Mori. 1999. On the evolution of hierarchical urban systems. *European Economic Review* 43: 209–251.

Hartwick, J. 2004. *Trade in a hierarchical system of cities. Mimeo*. Kingston: Department of Economics, Queen's University.

Krugman, P. 1995. *Development, geography, and economic theory*. Cambridge, MA: MIT Press.

Lösch, A. 1944. *The economics of location*. New Haven: Yale University Press. 1954.

McCann, P. 2001. *Urban and regional economics*. Oxford: Oxford University Press.

Starrett, D. 1978. Market allocations of location choice in a model with free mobility. *Journal of Economic Theory* 17: 21–37.

C

Central Planning

Tadeusz Kowalik

Central planning denotes the total body of government actions to determine and coordinate directions of national economic development. The process of central planning is composed of pre-plan studies and forecasts, formulation of aims for given periods of time, establishment of their priorities (order of importance), listing ways and means, and, eventually, the plan's implementation. Central planning is a term usually associated with Centrally Planned Economies (CPE) as opposed to Private Enterprise (or Market) and Mixed Economies (UN official classification), but it is often used in a broader sense to denote any systematic macroeconomic control by the government. For Tinbergen (1964), central planning means planning by governments, or national planning (in the Netherlands as well as in some other countries there are Central Planning Bureaux, even though these economies cannot be classed with the group of CPEs).

In this broader meaning, central planning takes several different names, specifically: 'direct', 'hierarchical' (Bauer 1978) or 'centralistic' as practised in most centrally planned economies; 'financial' as in Hungary; 'indicative' as in France.

The term 'planning' often stirs emotions. For some people, especially for many Communist economists, central planning is good by definition. Others use it to denounce socialism and indeed any kind of government intervention as 'planned chaos' (von Mises 1947). The scope

and meaning of central planning varies along with changing fashion. When Arthur Lewis confessed 'we are all planners now' (Lewis [1949] 1956, p. 74), it was fashionable to described any kind of state interventionism as 'planning'. Robbins (1947, p. 68) termed his proposal for a modest anti-inflationary or anti-deflationary fiscal policy as 'overall financial planning'. Since the 1970s, though, general opinion seems to have been increasingly wary of planning, indeed sceptical about its effectiveness. Accordingly, even some planners in the state administration who staunchly stood by that idea preferred to cover their activity under less emotionally charged terms (such as 'steering').

Initially, central planning used to be generally regarded as an inalienable feature of socialist economy and hence as the exact opposite of market and commodity production typical of capitalism. It was interpreted as planning in physical units, by central command, based upon a hierarchical structure of national economy which had at its disposal ways and means to enforce decisions by administrative order. Precisely this kind of planning system developed in the Soviet Union, less as a product of any definite concept or vision of socialist economy than as an outcome of many different interacting factors – doctrine and ideology, the specific situation of Russia at that time, and the political ends to which the victorious revolutionary authorities subordinated the economy.

Origins

After the Bolshevik victory in Russia Lenin's writings, apart from the above-mentioned view of planning as the exact opposite of market (which was shared by many other Marxists), provided two other theoretical contributions to the formidable task of organization of the economy. Following Rudolf Hilferding, Lenin (like Bukharin) described imperialism as an ante-chamber of socialism on account of the steadily accelerating process of production concentration (trusts) and the centralization of banks which were rapidly expanding their control of domestic industries. The German wartime economy with its large-scale combination of latest technology, planning and efficient organization, was viewed by Lenin as something like an archetype for a future socialist economy.

In the period of 'War Communism' (1918–20) the need for planning was repeatedly proclaimed but no national plan could actually be drawn up. It was only towards the end of the period that Gosplan, a planning commission, was created, although its job was modest and only vaguely defined for years thereafter. No firm way could be found to reconcile planning with the New Economic Policy (NEP) introduced in 1921.

The most important accomplishment of the early 1920s was the plan for electrifying all Russia, which was drawn up at Lenin's personal initiative in 1920 and which came to be referred to as GOELRO. That plan provided for the building, within the following 10–15 years, of power stations and related infrastructure in major industrial regions. At that stage, planning was viewed as primarily an engineering rather than economic activity (as can be seen if only from the composition of the commission, which included mostly engineers and agriculture specialists).

From 1925 onwards, Gosplan began to publish each year what were called economy-wide 'control figures' initially for a year only but later for five-year periods. Those figures were regarded as a non-binding set of estimations and forecasts. Their main contribution to the development of planning was that they eventually led up to the design of what is called the balancing method, which juxtaposes demand for goods with their output. First five-year plans also began to be drafted outside Gosplan.

The Soviet economy became a 'centrally planned' economy only at the time of the First five-Year Plan (1928/9–1932/3). That was a time of tough internal struggle in the party and one of escalating heroic development programmes. Each new draft version of the five-year plan, beginning with the first one after the Party Congress in December 1927 through to its final approval, set up increasingly ambitious tasks. But the balancing of tasks with resources in the plan was based mainly on overly optimistic (and largely unfeasible) forecasts of labour productivity growth. The party and the state authorities soon

began to mobilize the population to over-fulfil the plan, or, more precisely, those targets in the plan which were arbitrarily recognized as the most important ones (priority tasks). Thenceforward, plans became tools for mobilization rather than for balanced allocation of resources. Annual plans often shook up the current five-year plan to accommodate it to these new priorities (or super-priorities).

The First Five-Year Plan (which was officially declared fulfilled in four-and-a-quarter years) generated many bottlenecks and disproportions; this suggested that the pace should perhaps be slowed down – and priorities rearranged, as to some extent was attempted in the final version of the next five-year plan (for 1933–37). At the same time the new plan was even more detailed and its scope expanded significantly (the number of branches comprised by the plan increased to 120 from the original 50). The authors of the first five-year plans apparently did not realize the full institutional and political implications of over-ambitious tasks, the scale of which were in some cases downright unfeasible. In order to rescue those regarded as top priorities (especially those concerning heavy industries and manufacturing), others had to be sacrificed (those relating to standards of living were the first victims). This could only be accomplished by methods typical of wartime economy, that is, highly centralized organization, rigid subordination and discipline, all-embracing rationing, various kinds of coercion, and political mobilization. That was exactly what was attempted during the first two five-year plan periods.

To a considerable extent this amounted to a revival of the methods tried in the period of 'War Communism', including compulsory labour and rationing, however not as formal and lasting institutions like, for example, labour mobilization during the civil war, but either as side-effects of other campaigns (mass deportations during the collectivization drive, purges of the 1930s, etc), or as emergency responses to situations of extreme penury (rationing) which eventually should make room for allocation of labour and consumer goods through some kind of market (for ideological reasons the term was never used in relation to labour). This was combined with abandonment of the original egalitarianism in incomes policy; increased reliance on material incentives geared to plan fulfilment and piece rates became a distinctive mark of the Stalinist period.

C

Main Features (Formal Aspects)

The first two five-year plans set the general shape for a model of Soviet central planning, transplanted after World War II to communist Eastern Europe. That model survived unchanged through to the mid-1950s (except in Yugoslavia), and in most communist countries it functions to this day in its general outline.

In both its design and implementation stages, central planning is based on a hierarchical pattern of national economy, which in turn presupposes obedience and discipline. Freedom of choice (which is lifted only temporarily or partly) applies to purchases of consumer goods within the existing commodity supply and the state-determined purchasing power, as well as to choice of occupation and workplace within the statutory obligation to be in employment.

Using information on the economy's shape and tendencies at any given moment, the central authority formulates a set of general guidelines of the plan, possibly based on prior special studies and forecasts. The plan's guidelines include such aggregates as the distribution of the national income between accumulation and consumption, the shares and main directions of investment by sectors, the desired rate of overall economic growth etc. These guidelines as a rule are pre-defined by the leading bodies of the ruling party, and are then disaggregated by the government into guidelines for particular industrial ministries and local authorities to produce their own draft plans, which are further disaggregated and communicated to industrial associations and individual enterprises. Government guidelines include two kinds of indices; directives, which are mandatory for local planners in drafting their blueprints (whatever alteration may prove necessary can only be made by a superior agency) and information indices. The enterprise draft plans are

then aggregated by industrial associations and branch ministries, and their draft plans are in turn aggregated into a national (or central) economic plan for one or five years which is usually approved by parliament. Only after that are final corrections and adaptations introduced into lower-level plans. This particular procedure of plan construction has been called the 'spindle technique' in reference to textile machines, for guidelines and draft versions first travel from the top downwards, then up, and then again down the hierarchy.

One pivotal point in this procedure is the plan's internal consistency. The idea is to match demand for each particular resource with the level of its supply during the plan period. A whole system of balance sheets (indeed thousands of them) is used for that purpose. Balance sheets set – in physical or equivalent units – available amounts of materials, capacities, energy, labour, as well as financial means (personal income and spending, foreign trade balance, the budget) against anticipated demand in each case.

Plan fulfilment is a fundamental obligation of each economic organization. Managers and, to some extent the workforce as well, are evaluated for their plan performance and rewarded or penalized accordingly. Tasks named in an enterprise plan are both commands by a superior authority and obligations to supply enough resources to safeguard smooth cooperation. Although enterprises are given not only quantitative targets but also qualitative ones (e.g. technological input/output coefficients for materials, power etc, the importance of output–quantity performance is overriding.

Advantages and Failures

This particular model of planning was conceived in a country with abundant resources of labour (open or disguised unemployment) and primary products; it was applied also in several other countries with large unused capacities. Providing able to mobilize idle resources initially produced very high growth rates, although one cannot take official statistical records at their face value. Determination of obligatory priorities on a national scale enabled countries to concentrate resources and efforts on several selected spectacular tasks. The successful bid to transform the Soviet Union into a superpower in a relatively brief time is perhaps the least debatable success of this planning model.

However, from the mid-1950s onwards centrally planned economies have been coming under growing criticism both from professional economists and from the general public. The criticism became particularly sharp as growth indicators declined and started to affect the (slow anyway) improvement of living standards; the system's weakness in generating and absorbing technological innovations became increasingly evident. However, critical voices – even when acknowledged by political authorities – did not lead, as a rule, to consistent and effective changes in the economic system.

The main lines of criticism of deficiencies of the existing system of central planning can be summarized as follows:

The procedure for building plans outlined above cannot guarantee efficient allocation of resources. The tasks and resources for their implementation are not decided by the central planning agency in a truly 'sovereign' way because such an agency is bound to rely on the supply of information from lower-rank agencies. But that information, apart from some natural delays or mistakes made in its transmission upwards, is often deliberately distorted by enterprises, which use it as a weapon in plan bargaining. Enterprises usually want to wrench as large means and as small tasks as possible from the central economic authority for themselves. Industrial association, indeed even branch ministries, often helps them achieve this purpose. At that stage, too, the main battle for investment funds begins. Enterprises and local authorities try to get 'put on' the plan by deliberately underrating estimated costs of their undertakings. Eventually, the plan is apparently brought into balance, but it has built-in significant disproportions right from its start, which leads to a waste of resources.

Even greater waste results from the centralistic bureaucratic method of controlling the execution of plan tasks, which eventually leads to equating planning with management. The over-taut plan,

based as it is on unrealistic assumptions, especially regarding labour productivity growth, can later be 'fulfilled' only by setting up a whole system of ad hoc priorities and superpriorities which makes a reduction of nonpriorities unavoidable. As a rule, the victimized sectors are those related to the sphere of personal incomes or social or municipal services (public transport), the health service, housing, education – treated as residuum.

Once they have been assigned the required resources by the central economic authority, enterprises no longer feel compelled to seek ways of saving materials or energy. Because deliveries of materials and energy are as a rule irregular, enterprises try to provide against such risks by hoarding excessive inventories of materials and reducing employment only reluctantly. Moreover, enterprises are given no effective inducements to seek new technology, indeed even to emulate existing new techniques.

Prices set by the central authority are as a rule rigid and random, reflecting neither costs nor relative scarcities of individual goods. As a consequence, both at the central level and at enterprise level clear criteria of choice are largely absent.

Viewed from the consumer's vantage point, centrally planned economies provide poor-quality goods and a meagre product mix. Their incapability of meeting greater diversification of needs, which inevitably progresses along with increase in income levels, is one of the major reasons for the growth of a 'second economy' (moonlighting, corruption etc).

Over-taut plans, implemented through commands, unavoidably generate an inflated control system and subject the economy to political goals. Subordination of economies to politics is often presented as expression of general (social) interest; in reality this subordination often conceals vested interests of small informal groups. In the process of plan negotiations and rearrangement of priorities in the course of implementation, centralistic administrative planning engenders informal lobbies which exert growing pressure on the central authority. A product of quasi-missionary zeal to develop the production of means of production, the heavy industry lobbies are the strongest of all. Gradually, the central authority is losing its 'sovereignty' to them. Even when the authority begins to appreciate 'harmony' more than 'rush' (Kornai 1972) it is unable to shed that pressure.

This very role of lobbies goes against the widely held belief that in the centrally planned economies the superior position belongs to the preferences of the central planners. Increasingly concrete decisions are made under growing pressures of various informal vested interest groups. In this situation, criteria of choice cannot be clear or unequivocal, which makes public control of the central planning agency's operations even more difficult. For the same reason, and even more because of the secretive style of work of state agencies, as well as absence or limitation of consumer organizations, environmental groups, independent trade unions, and with restricted press freedom, the central authorities cannot play the part of an umpire reconciling different social interests. Protection of public interest becomes fictitious under these conditions. Thus, when official doctrines proclaim unity of interests, this may simply conceal a growing tendency towards a peculiar kind of 're-privatization' of centrally planned economics.

Evolution and Prospects

Since the mid-1950s, in the system of central planning as practised in countries of the so-called 'real socialism' two categories of change have taken place.

The growth and mathematization of economics, in particular the expansion of linear programming, operations research, input–output analysis, cybernetics and systems analysis, the wide extension of computer applications etc., have supplied planners with subtler tools for their work. The development of these tools fuelled hopes, already in the 1960s, that planning would proceed 'from balancing the plan towards the choice of optimal plan' (Lange 1965). 'Planometrics' came into use then, indeed even something like a 'computopia' began to develop.

The second kind of change was more institutional in character. It came along with de-Stalinization, of which economic reform was and still remains a part. Unlike in Yugoslavia,

where the economic system was to correspond to an entirely different model of socialist society compared with the Soviet-type one, in the countries belonging to CMEA institutional changes amounted, generally speaking, to a transfer of some economic decision-making to lower-level units, an expansion of material incentives for managers and workers alike, and an extension of market mechanisms.

As a result of the new techniques and of the partial decentralization, central planning has probably become a slightly more efficient tool of economy-wide control. However, all those improvements were ultimately too negligible and inconsistent to stand up to the growing complexity of economy, in particular to offset the depleting reserves of extensive-type growth factors (excess labour, cheap raw materials) by more intensive methods of growth stimulation. The technology gap between CMEA and advanced Western countries, which became clear in the 1960s and has kept widening since then, has not been bridged; if anything it has continued to widen. Hence, repeated calls for more or less radical economic reform are still the order of the day.

Planning and Freedom

Ever since its inception, the question of economic planning has set off disputes about democracy and individual freedom. In its original purely ideological concept, planning used either to be equated with democracy or presented as democracy's exact opposite: suffice it to mention the New Leftist utopia of a social system based on the belief that production and distribution can somehow be planned by the people with a total absence of market and state. The eternal Kingdom of Freedom was to come simply as soon as market and state alike have been abolished.

More elegant, albeit no less utopian, is the free-marketeers' blueprint for rejecting any governmental planning as a threat to efficiency and freedom. Although quite fashionable (and not only in the West), this mode of thinking is nonetheless outside the mainstream of disputes over planning versus freedom.

In fact, most major currents of social thinking have undergone a process of radical rethinking in the course of recent decades. This holds for liberalism (Mannheim 1940; Galbraith 1973; Lindblom 1977) and for non-Communist socialism (Crosland 1956; Crossman 1965; Nove 1983) as well as for Marxism (Brus 1975; Horvat 1982; Kornai 1985). Whatever differences may divide all these currents of thought, as indeed individual thinkers within each current, all of them are aware of two kinds of threat to freedom – one that comes from allembracing, hierarchical and bureaucratic planning, and another that comes from the failure to plan anything at all. The market mechanism is regarded as something like a barrier to bureaucratic arbitrariness. But its failures in turn may put at hazard not only economic but even political stability, thereby destroying the foundations of the desired social order. Planning, within given limits, thus turns out to be an indispensable condition of freedom. While making a plea for a polycentric model of economy – both in the sense of providing for different forms of ownership and of decision-making – all these currents of thinking believe that society as a whole should have an authentic say (via its representatives) on the main lines of investment and on general rules for national income distribution.

Of course, there is nothing inevitable in the long-run direction this movement will take either in the West or in the East. The chance to create a social order which would be based upon the three main tiers of plan, the market and freedom would be much greater if it were clear that each of these is a necessary condition for high socio-economic efficiency, and that freedom too can be viewed not only as a value in itself but also as a specific kind of production factor. Some authors have questioned this dependence of economic efficiency on political democracy (Gomulka 1977). However, neither studies of this relationship in many Third World countries (Adelman and Taft 1967) nor the record of previous reforms in the Communist world supply any definite answer to this question. On the other hand, the analysis of pressures on, and prospects of, the evolution of Communist systems in Eastern Europe has led to a rather persuasive argument (Brus 1980) that

without democratizing internal political relations these systems will be unable to remove (or at least to reduce substantially) central planning's chronic deficiencies, such as insufficient and distorted information flows, negative selection of managerial personnel, chronic investment failures, labour alienation etc. The stagnation threatening the Communist countries presses the ruling groups to more radical reforms which would combine plan, market and freedom. At the same time, repeated setbacks of neoliberal economic policies in the West may well generate fresh and strong public pressure for changes in a similar direction.

See Also

- ▶ Command Economy
- ▶ Decentralization
- ▶ Market Socialism
- ▶ Material Balances
- ▶ Socialism

Bibliography

Adelman, I., and C.M. Taft. 1967. *Society, politics and economic development: A quantitative approach*. Baltimore: Johns Hopkins Press.

Bauer, T. 1978. Investment cycles in planned economies. *Acta Oeconomica* 21(3): 243–260.

Brus, W. 1975. *Socialist ownership and political systems*. London: Routledge & Kegan Paul.

Brus, W. 1980. Political system and economic efficiency: The East European context. *Journal of Comparative Economics* 4(1): 40–55.

Cave, M., and P. Hare. 1981. *Alternative approaches to economic planning*. New York: St Martin's Press.

Crosland, C.A.R. 1956. *The future of socialism*. London: Jonathan Cape.

Crossman, R.H.S. 1965. Planning and freedom. In *Essays in socialism*, ed. R.H.S. Crossman. London: Hamish Hamilton.

Davies, R.W., and E.H. Carr. 1974. *Foundations of a planned economy 1926–1929*. Harmondsworth: Penguin.

Ellman, M. 1983. Changing views on central economic planning: 1958–1983. *The ACES Bulletin, A Publication of the Association for Comparative Economic Studies* (Tempo, Arizona) 25(1).

Galbraith, J.K. 1973. *Economics and the public purpose*. Boston: Houghton Mifflin.

Gomulka, S. 1977. Economic factors in the democratization of socialism and the socialization of capitalism. *Journal of Comparative Economics* 1(4): 389–406.

Horvat, B. 1982. *The political economy of socialism. A marxist social theory*. Armonk: M.E. Sharpe.

Kornai, J. 1972. *Rush versus harmonic growth*. Amsterdam: North-Holland.

Kornai, J. 1985. *Contradictions and dilemmas, studies on the socialist economy and society*. Corvina: Kner Printing House.

Lange, O. 1965. Od bilansowania do wyboru optymalnego planu (From balancing the plan to the choice of optimal plan). *Nowe Drogi* (Warsaw) 2.

Lewis, W.A. 1949. *The principles of economic planning*. London: George Allen & Unwin Ltd, 1956.

Lindblom, C. 1977. *Politics and markets. The world's political-economic systems*. New York: Basic Books.

Mannheim, K. 1940. *Man and society in an age of reconstruction. Studies in modern social structure*. London: Routledge & Kegan Paul, 1974.

Nove, A. 1983. *The economics of feasible socialism*. London: Allen & Unwin.

Robbins, L. 1947. *The economic problem in peace and war*. London: Macmillan.

Tinbergen, J. 1964. *Central planning*. New Haven/London: Yale University Press.

von Mises, L. 1947. *Planned chaos*. Irvington-on-Hudson: The Foundation for Economic Education.

Centre of Gravitation

Luciano Boggio

In the Classical theory concerning the price of commodities there are two notions of price: the *market price*, which is the price actually prevailing, and the *natural price*, which is equal to 'what is sufficient to pay the rent of the land, the wages of the labour, and the profits of the stock. . .according to their natural rates' (Smith 1976, p. 72); and the natural price is a *centre of gravitation* for the actual price, i.e. the latter is continually tending to the former.

In the description of this tendency given by the Classics, one can distinguish two main propositions. According to the first one, the market price depends on the difference between current supply and 'effectual demand' – which is 'the demand of those who are willing to pay the natural price of

the commodity' (Smith 1776, p. 73) – the market price being higher, lower than, or equal to the natural price, if such a difference is, respectively, negative, positive or zero. According to the second proposition, the difference between market price and natural price gives rise to movements of capitals and changes in the structure of production, so that the output of a commodity increases (decreases) if such a difference is positive (negative).

The position of the economy in which market price equals natural price and output equals effectual demand is a 'centre of repose and continuance' (Smith 1776, p. 75), a position which is bound to repeat itself unaltered, until an exogenous change (e.g. in the available techniques of production) takes place.

On the basis of current definitions and methods of dynamical analysis (see, e.g. Lasalle 1976), one can recognise in the above theory the description of a dynamical system, in which the state variables are prices and output (or capital) quantities and for which the vector formed by natural prices and the corresponding output (or capital) quantities is an equilibrium.

The notion of a uniform rate of profit price vector, towards which actual prices tend to move, having been accepted – after Ricardo – by generations of economists, including Marshall and Walras, was then abandoned in the works on general equilibrium of recent decades.

The uniform-rate-of-profit price vector – let us call it *production price vector* – had a central place in Sraffa's book *Production of Commodities by Means of Commodities, Prelude to a Critique of Economic Theory* (1960), but nothing was said in it about the relation between such price vector and actual prices.

As a 'prelude to a critique of economic theory' Sraffa's analysis does not require such a relation; but it does, if it is to be used as a building block for a 'reconstruction' of economic theory. For this reason the Classical theory, according to which 'long period positions' are 'centres of gravitation' for actual prices, has been recently reintroduced (Garegnani 1976).

This reintroduction, however, also requires a critical re-examination today in the light of the contemporary methodology of dynamics. In order to illustrate this point, we shall try to formalise the gravitation theory of the Classics in a simple way.

Let us consider an economy in which there are n sectors, each producing a different commodity. The production of every commodity requires capital advances, which must cover the purchases of material inputs as well as the payment of a subsistence wage rate to each employed worker.

To simplify the discussion further, we also *assume* that for our economy a unique semi-positive price vector exists, such that in each sector the following two conditions hold together:

(a) a uniform rate of profit prevails;
(b) demand and supply are equal.

We call this vector 'the production price vector' of our economy and denote it by $p^* = p_i^*$, $i = 1, 2, 3, \ldots, n$. The existence of such a vector can be *proved* in various kinds of models (see, e.g. Boggio 1985).

Let us now reconsider the two propositions in which we have summarised the description of the gravitation process given by the Classics and call d_{it}, p_{it} and q_{it}, respectively, the effectual demand for, the actual price and the current output of the i-th commodity at time t, $t \in R_+$.

Then a simple way to model the first proposition is the following

$$p_{it} - p_i^* = g_i(d_{it} - q_{it}), \qquad i = 1, 2, \ldots, n \quad (1)$$

where g_i is a continuous sign-preserving function.

As for the second proposition, that is, the relation between output changes (as determined by capital movements across sectors) and profitability (as expressed by the difference between market and natural price), it can be modelled as

$$\dot{q}_i = s_i\left(p_{it} - p_i^*\right), \qquad i = 1, 2, \ldots, n\frac{1}{2} \quad (2)$$

where $\dot{q}_i = dq_{it}/dt$ and s_i is a continuous sign-preserving function.

These two equations, however, are not sufficient to prove the gravitation thesis. They must be

supplemented by some specific assumption about the time-pattern of d_t, the vector of effectual demands. Let us assume that d_t is constant

$$d_t = d^* \tag{3}$$

where d^* is a fixed semi-positive vector. Then from (1), (2) and (3) we get

$$\dot{q}_i = s_i\left[g_i\left(d_i^* - q_{it}\right)\right] \tag{4}$$

Since $d\left(q_{it} - d_i^*\right)$ is always equal to $\dot{q}_i$, one can see that, by equation (4), its sign is always opposite to that of $\left(q_{it} - d_i^*\right)$. Therefore as t grows $\left|q_{it} - d_i^*\right|$ decreases motonically, i.e. q_{it} tends to d_i^*. This implies that as t tends to d_i^*, must also, by equation (1), tend to d_i^*.

Hence the couple (p^*, d^*) is a globally asymptotically stable equilibrium for system (1)–(2)–(3). Notice that the same conclusion could be reached if we assumed that d_{it}, instead of being constant, were growing at a constant proportional rate.

This result means that, according to our model, although exogenous changes may shift the actual price vector and/or the production price vector, the gravitation mechanism will always tend to close the gap between them.

However, our model and the gravitation theory of the Classics are rather crude and, in several points, unsatisfactory. A clear example of this is the assumed equivalence between a positive (negative) value of $\left(p_{it} - p_i^*\right)$ and a capital-attracting (-repelling) sectoral rate of profit. If the higher than average rates of profit are capital-attracting and the lower than average rates are capital-repelling, one can show that such an equivalence does not hold in general (see Steedman 1984). The ratio between a given sectoral profit rate and the average rate depends also on the prices of the commodities required as input in that sector. The Classical description neglects these inter-industry links and any attempt to incorporate them in the analysis requires the simultaneous consideration of all prices, output levels, etc. thereby disrupting the beautiful simplicity of the Classical theory of gravitation.

The study of the stability of production prices by means of the methods of contemporary dynamic analysis has recently begun (for references and more comments, see Boggio 1985). Two main approaches have been followed. In the first approach the process of price formation is based on some kind of mark-up or full-cost rule. Very strong stability results are obtained. But the assumption of exogenous mark-ups or target profit rates is not entirely satisfactory. In the second approach the original description of the Classics is followed more closely: actual price changes depend upon supply and demand and output changes depend upon profit differentials.

We notice that in the latter approach the equilibrium vector is formed by both production (relative) price vector and steady state output proportions. Since no economy grows in a balanced way, the reference to balanced growth is often considered such a weakness as to deprive a theory of any usefulness. Actually, the meaning of balanced growth, as of every equilibrium concept, is not necessarily to offer in itself a description of reality. If the equilibrium is stable, the effects of changes in the data of the system can be approximately studied by means of the displacements of equilibrium positions: in the case under discussion, the effects on prices by the displacements of production prices, the effects on output proportions by the displacements of balanced growth proportions. A condition for the correct use of this method, an outstanding example of which in the field of economics is the above described theory of the Classics, is that the changes in the data should be slower than the movements of the state variables of the (dynamical) system. These remarks suggest the great advantage of studying change by means of comparisons or sequences of equilibria: the more fundamental aspects of change are selected out of the variety of accidental and transitory ones.

As for the results obtained in the latter approach to the study of the stability of production prices, they are mainly against stability, except for the case when strong price substitution effects in consumption are introduced. A more promising approach, in terms both of realism and of stability results, consists, probably, in assuming that price

changes depend *mainly* on cost changes, but some role in determining the former is also played by excess demand.

Much more work in this field, however, seems necessary. Its importance derives not only from the question of 'gravitation' itself, but also from more general issues of economic theory.

By specifying in a rigorous and formal way a dynamical process for which Sraffa prices are (part of) an equilibrium and by showing that such a process is not reducible to a neo-Walrasian disequilibrium process – in which prices react to excess demands, supply *instantaneously* adjusts to prices and expected future prices are replaced by current prices of 'futures' – such a work can establish in the clearest way that Sraffa prices are not simply reducible to a special case of neo-Walrasian general equilibrium theory.

Secondly, if it can show that, to give a stylized description of price-quantity dynamical interrelationships, there are more plausible ways than the neo-Walrasian one, such a work can give a contribution in the direction of replacing the neo-Walrasian paradigm.

See Also

- ▶ Competition
- ▶ Equilibrium (Development of the Concept)
- ▶ Long Run and Short Run
- ▶ Market Price
- ▶ Natural and Normal Conditions

Bibliography

Boggio, L. 1985. On the stability of production prices. *Metroeconomica* 37(3): 241–267.

Garegnani, P. 1976. On a change in the notion of equilibrium in recent work on value and distribution; a comment on Samuelson. In *Essays in modern capital theory*, ed. M. Brown, K. Sato, and P. Zarembka. Amsterdam: North-Holland.

Lasalle, J.P. 1976. *The stability of dynamical systems*. Philadelphia: SIAM.

Ricardo, D. 1817. In *On the principles of political economy and taxation*, ed. P. Sraffa. Cambridge: Cambridge University Press, 1951.

Smith, A. 1776. In *An inquiry into the nature and causes of the wealth of nations*, ed. R.H. Campbell and A.S. Skinner. Oxford: Oxford University Press, 1976.

Sraffa, P. 1960. *Production of commodities by means of commodities: Prelude to a critique of economic theory*. Cambridge: Cambridge University Press.

Steedman, I. 1984. Natural prices, differential profit rates and the classical competitive process. *The Manchester School of Economic and Social Studies* 52(2): 123–140.

Certainty Equivalence

Xavier Freixas

JEL Classifications
D8

In order to take a decision in an uncertainty context, it is necessary, from a theoretical point of view, to build a model and specify all the consequences in every possible state of the world. In applied work this method is much too involved.

Consequently, for applied purposes, it would be interesting to have a model where uncertainty is treated in such a way that the decision problems are as simple as the equivalent ones in a certainty framework. The identification of the conditions under which such an isomorphism between the optimal decisions under uncertainty and the optimal decisions in an equivalent certainty context holds is called the certainty equivalent problem.

Theil (1954) has been the first to point out the problem and to suggest a specific model in which the certainty equivalent property holds.

Theil imposes the following two assumptions: (i) the vector x of instruments and the vectory y of result variables are related by a simple equation

$$y = g(x) + S \tag{1}$$

where S is a vector of random variables, that we can take to have a zero expected value without loss of generality. (ii) The decision-maker's objective function is quadratic and can be written as

$$u(x,y) = A(x) + \sum_{i=1}^{m} A_i(x)y_i + \frac{1}{2}\sum_{i=1}^{m} \times \sum_{y=1}^{m} A_{ij}y_iy_j \qquad (2)$$

Using such a model it is straightforward to show that whenever the optimal solution to the problem of maximizing the expected utility under the constraint (1) exists, it is the same as the optimal solution to the equivalent certain problem:

$$\begin{cases} \text{Max } u(x,y) \\ y = g(x) \end{cases}$$

This result is extended not only to the multi-period problem but also to the case where the decision-maker receives more and more information as time elapses. The resulting stochastic problem is then more involved, but it is simply solved by use of dynamic programming, the optimal strategy in period t being a function of the previously observed signals η_{t-k}

$$X_t^* = x_t^*(\eta_1, \eta_2, ..., \eta_{t-1})$$

Again, the conditions for the first period solution to this problem to be the solution of the equivalent certain problem are very strong. As before, it has to be the case that the objective function is quadratic, but in addition the constraint relating instruments to results is restricted to be of the following type:

$$y = RX + S$$

where R is a matrix with some required specifications (namely, the value of the instrument variables of one period have no effect on the result variables of the preceding periods).

The conditions that guarantee the equivalence between the uncertainty problem and the certainty problem are so restrictive, that an alternative view of the problem has been suggested. Instead of setting restrictions on the parameters of the model, the uncertainty itself is restricted to be 'small'. Formally, this is equivalent to consider an entire class of problems that can be ranked in their uncertainty as measured by a parameter ε and whose limit is the certain problem. The question is then to know under what conditions the solution to the limit of the random problems, that is equal to the one of the certain problem, is independent of ε to the first order, so that

$$\frac{dE\left[x_t^*(\eta_1, ..., \eta_t)\varepsilon\right]}{d\varepsilon} = 0 \qquad \text{for } \varepsilon = 0.$$

This slightly different point of view is called the 'first order certainty equivalence' problem and has been dealt with by Theil (1957) and Malinvaud(1969).

The very general conditions obtained by Malinvaud for the first order certainty equivalent to hold are (i) that the objective function is twice differentiable and (ii) that the optimal strategy is continuous with respect to the degree of uncertainty. If this condition holds, the optimal values of the instruments at time 1 are, to the first order approximation, independent of the degree of uncertainty.

It is clear that this condition cannot be met if there are constraints on the future instrument variables, since this will bring in a kink. A particular and natural example of a framework where the first order certainty equivalence does not hold is when decisions are irreversible. As pointed out in Henry (1974), it is then the case that the value of the decision in the first period will affect the decision set in the following periods, and consequently, the use of the certainty equivalent would generate a systematic error.

See Also

▶ Risk

Bibliography

Henry, C. 1974. Investment decisions under uncertainty: The irreversible effect. *American Economic Review* 64: 1996–2012.

Malinvaud, E. 1969. First order certainty equivalence. *Econometrica* 37: 706–718.

C

Simon, H. 1956. Dynamic programming under uncertainty with a quadratic criterion function. *Econometrica* 24: 74–81.
Theil, H. 1954. Econometric models and welfare maximization. *Weltwirtschaftigsches Archiv* 72: 60–83.
Theil, H. 1957. A note on certainty equivalence in dynamic planning. *Econometrica* 25: 346–349.

CES Production Function

Ryuzo Sato

Abstract

The CES (constant elasticity of substitution) production function, including its special case the Cobb–Douglas form, is perhaps the most frequently employed function in modern economic analysis. Not only is the CES function used for the formal depiction of production technology, it is used as a convenient tool for empirical analysis as well. In addition to production theory, the CES function, more commonly known as the Bergson family of utility functions, is employed in utility theory.

Keywords

Bergson family of utility functions; CES production function; Cobb–Douglas functions; Elasticity of substitution; Jensen inequalities

JEL Classifications

E23

The CES (constant elasticity of substitution) production function, including its special case the Cobb–Douglas form, is perhaps the most frequently employed function in modern economic analysis. Not only is the CES function used for the formal depiction of production technology, it is used as a convenient tool for empirical analysis as well. In addition to production theory, the CES function, more commonly known as the Bergson family of utility functions, is employed in utility theory.

Ordinary CES Production Functions

The simplest form of CES function utilized in production theory is the constant returns to scale type (Arrow et al. 1961):

$$Y = T[\alpha K^{-\rho} + (1-\alpha)L^{-\rho}]^{-(1/\rho)} \qquad (1)$$

where $Y =$ output, $K =$ capital, $L =$ labour, and the parameters T, α and ρ satisfy the conditions: $T \geq 0$, $0 \leq \alpha \leq 1$ and $\rho \leq -1$. As is implied by its name, the elasticity of factor substitution between capital and labour for production function (1) is expressed as some constant value.

For any neoclassical production function, $Y = f(K, L)$, the elasticity of factor substitution between capital and labour is defined as the proportionate change in the K/L ratio (k) relative to the proportionate change in the marginal rate of factor substitution $r = f_L/f_k$ along a given isoquant curve, where $f_L = \partial Y/\partial L$ and $f_K = \partial Y/\partial K$ are the respective marginal products. That is,

$$\sigma = \frac{\mathrm{d}\log k}{\mathrm{d}\log r} = \frac{f_k f_L (f_k K + f_L L)}{KL\left(2f_{KL}f_K f_L - f_K^2 f_{LL} - f_k^2 f_{KK}\right)}, \qquad (2)$$

where σ represents the elasticity of substitution and f_{KL}, f_{KK} and represent the cross and own derivatives of the respective marginal products.

Applying definition (2) to production function (1) we obtain:

$$\sigma = \frac{1}{1+\rho} \quad \text{or} \quad \rho = \frac{1-\sigma}{\sigma}. \qquad (3)$$

Consequently, it is easy to see why ρ is often referred to as the 'substitution' parameter. The α parameter in production function (1) is the 'distribution' parameter that permits the relative importance of capital and labour to vary in production. In the extreme case where $\rho \to 0$ or $\sigma = 1$ the CES function (1) converges to the Cobb–Douglas form:

$$Y = TK^{\alpha}L^{1-\alpha}. \qquad (4)$$

In this form, it is evident that α and $1-\alpha$ are the production elasticities of capital and labour respectively. Under conditions of perfect competition, α and $1-\alpha$ will also equal the respective relative income shares (or income distribution). The T parameter in both production functions (1) and (4) is the 'efficiency' (or technical progress) parameter.

With the exception of its special case the Cobb–Douglas form, the ordinary CES production function is cumbersome and difficult to manipulate. However, the underlying expression for the marginal rate of factor (technical) substitution has a simple form and this is the primary reason for the popularity and wide use of this production function.

Homothetic and Non-homothetic CES Production Functions

Any monotonic transformation of the ordinary CES production functions (1) belongs to a class of CES production functions called the homothetic class, that is,

$$Y = F(f), F' > 0,$$

where

$$f = T[\beta K^{-\rho} + (1-\beta)L^{-\rho}]^{-(1/\rho)}. \tag{5}$$

In addition to the class of homothetic CES production functions, there is a more general, and perhaps more meaningful, class of non-homothetic CES production functions. One can refer to the class of non-homothetic CES functions as the 'general class' of CES production functions as it contains the homothetic class as a special case.

The class of non-homothetic CES production functions is derived as a solution to the differential equation that defines a constant elasticity of factor substitution. However, unlike the case of the homothetic CES production functions where the marginal rate of factor substitution is (implicitly) assumed to be independent of either the output level and the process of technical change, the family of non-homothetic CES production functions explicitly assumes that output level and technical change will have some kind of impact on the factor input ratio.

The class of non-homothetic CES production functions can be expressed as follows (Sato 1975):

$$C_1(Y)K^{-\rho} + C_2(Y)L^{-\rho} = 1, \quad \rho = \frac{1-\sigma}{\sigma}, \sigma \neq 1, \tag{6a}$$

$$C_1(Y)\log K + C_2(Y)\log L = 1, \sigma = 1, \tag{6b}$$

where C_1 and C_2 are functions of the output level Y. When $C_1 = aC_2$, where a is a constant, we can express (6a) as

$$K^{-\rho} + aL^{-\rho} = \frac{1}{C_1(Y)} = B(Y)$$

or

$$Y = B^{-1}(K^{-\rho} + aL^{-\rho}).$$

Note that with the appropriate choice of B and a, we can always express the above in the form of the ordinary CES production function. In general, the non-homothetic CES production functions are in an implicit form and can never be expressed in an explicit form.

Classification of Non-homothetic CES Production Functions

The general class of non-homothetic CES production functions can be classified in a number of ways, depending on the specific purpose in mind. For example, it is well known that the ordinary CES production function belongs to the explicit and separable class of homothetic CES functions. In a similar fashion, we can derive an explicit and separable class of non-homothetic CES functions (Sato 1974). Another way of

classifying non-homothetic CES production functions is to consider the form of the underlying marginal rate of factor substitution function. However, the most precise way of classifying the family of non-homothetic CES production functions is to utilize Lie group theory.

A Historical Note

It was Arrow et al. (1961) who first utilized the ordinary CES production function expressed in (1) for the estimation of constant returns to scale aggregate production functions using cross-country data. Since then, the ordinary CES function and its variants have been widely applied in both theoretical and empirical work involving production behaviour.

Prior to its application to production analysis, the ordinary CES function, was utilized in the study of demand as the Bergson family of utility functions (Samuelson 1965). Earlier writers in growth economics, such as Dickinson (1955) and Solow (1956), used special cases of the CES function, such as $\sigma = 2$. In the field of mathematics, Courant (1959, vol. 1, pp. 557, 601) has used the explicit form of the ordinary CES function in conjunction with the so-called Jensen inequalities.

A published note by McElroy (1967) contains the first reference to the non- homothetic CES production family. However, it was not until later that Sato (1974) derived an explicit form of the non-homothetic CES production function. The application of Lie group theory to CES production functions was first presented in 1975. This work demonstrated that the 'projective' type of technical change with eight essential parameters can be used most effectively to classify the general non- homothetic CES family of production functions. This work is summarized in Sato (1981, ch. 5).

See Also

- ▶ Cobb–Douglas Functions
- ▶ Elasticity of Substitution

Bibliography

Arrow, K., H. Chenery, B. Minhas, and R. Solow. 1961. Capital–labour substitution and economic efficiency. *Review of Economics and Statistics* 43: 225–250.

Courant, R. 1959. *Differential and integral calculus*, 2 vols. New York: Wiley.

Dickinson, H. 1955. A note on dynamic economics. *Review of Economic Studies* 22 (3): 169–179.

McElroy, F. 1967. Note on the CES production function. *Econometrica* 35: 154–156.

Samuelson, P. 1965. Using full duality to show that simultaneously additive direct and indirect utilities implies unitary price elasticity of demand. *Econometrica* 33: 781–796.

Sato, R. 1974. On the class of separable non-homothetic CES production functions. *Economic Studies Quarterly* 25 (1): 42–55.

Sato, R. 1975. The most general class of CES functions. *Econometrica* 43: 999–1003.

Sato, R. 1981. *Theory of technical change and economic invariance*. New York: Academic Press.

Solow, R. 1956. A contribution to the theory of economic growth. *Quarterly Journals of Economics* 70: 65–94.

Ceteris Paribus

John K. Whitaker

Keywords

Ceteris paribus; Endogeneity and exogeneity; Partial equilibrium; Time-period analysis

JEL Classifications

B4

The Latin phrase 'ceteris paribus', which translates as 'other things the same', is much invoked by economists. Its popularity stems from its prominent use by Alfred Marshall (1920, pp. xiv–xv, 366–70), who invented the metaphor of 'the pound called Coeteris Paribus' – pound being used here in the same sense as in impoundment – in which are imprisoned 'those disturbing causes, whose wanderings happen to be inconvenient' (Marshall 1920, p. 366).

The term 'ceteris paribus' has no clearly settled technical meaning among economists, so that an attempt to chronicle its usage would be both difficult and unrewarding. Instead, it seems preferable to distinguish the most important alternative ways in which the phrase might be employed, alluding only briefly to the pertinent literature. It is important to distinguish at the outset three broad ways in which the phrase might be used. These are:

- As a reminder that any practicable theory must take for granted the stability and continuance of certain background circumstances;
- As a warning, when using a theory predictively, that certain variations in circumstances admitted by the theory have been assumed not to occur;
- As an instruction to hold hypothetically constant some members of a set of *necessarily* covarying variables while changes in the others are contemplated.

For example, an analysis of the movement of a group of adjacent cooling towers during gales might (i) abstract from earthquakes, or (ii) hold constant ambient temperature while considering the effects of varying wind speed, or (iii) analyse the swaying of one tower in a high wind on the assumption that the other towers are perfectly rigid, even though they too must actually sway in a way that subtly alters the wind currents buffeting the first tower. In the language of econometric models, these three usages of 'ceteris paribus' can be characterized as (i) a reminder that the model's structure is assumed not to change, or (ii) a warning that certain *exogenous* variables are presumed to remain constant when others change, or (iii) an instruction to hold constant certain *endogenous* variables while varying others, even though this is not justified by any separability properties of the model's structure.

The first two usages pose no difficulties. In each, the invocation of ceteris paribus merely serves as a reminder that a more comprehensive or elaborate analysis might have been attempted. The risk of earthquakes could have been incorporated into the analysis of cooling-tower stability at the price of added complexity. But a failure to do so is without methodological significance. The incidence of earthquakes is unlikely to be affected by any movement of the towers, so that the exclusion merely singles out a convenient stopping place on the inevitable trade-off between comprehensiveness and complexity. Analogously, in predicting with an econometric model it would be possible to make careful predictions of the changes in all exogenous variables that accompany a tax cut. But a failure to do so involves no logical inconsistency, and the resulting ceteris-paribus prediction of the tax cut's effects will still have substantive interest.

It is the third usage alone, with its implied logical inconsistency, which poses distinct difficulties of interpretation and methodological justification. To start with, the assertion that certain variables are mutually interdependent presumes knowledge, at least in principle, of a correct comprehensive theory in which these variables are endogenous. For economists, the requisite background theory has usually been that of Walrasian competitive general equilibrium. In such a context, the invocation of ceteris paribus in its third sense to freeze hypothetically certain endogenous variables (or, more generally, to treat them as if exogenous) can itself be given at least three alternative rationalizations.

Partial Equilibrium Analysis as an Approximation

The focus here is on the demand-supply interactions in one market or a few closely interrelated markets as exogenous shifts occur, prices in all other markets being treated as hypothetically constant (or perhaps in some cases varied exogenously). Such a procedure is inconsistent with the supposed background general-equilibrium theory which implies that all prices vary interdependently. But it may give an adequate *approximate* representation of the particular markets being examined (see Viner 1953, p. 199). This is more likely the weaker and more diffuse are connections to, and feedbacks from, markets outside the examined set. Smallness relative to the

entire economy is usually helpful in this regard, but such questions have received surprisingly little detailed analysis.

Approach by Successive Approximation

Here the use of ceteris paribus restrictions is viewed as a necessary transitional step towards the evolution or understanding of a fully-comprehensive general-equilibrium theory. The limitations of human comprehension, its need to understand and test only one link of a complete chain at a time, calls for a piecemeal step-by-step progression from the crude but simple to the sophisticated but more complex, even though such a proceeding would appear illogical to an all-comprehending Cartesian intelligence. It should, however, be observed that this progression could well take place by starting with a highly aggregated general equilibrium model and successively reducing the degree of aggregation, instead of by starting with a simple partial-equilibrium model and gradually expanding its coverage until general equilibrium is reached – as is Marshall's clearly stated strategy (Marshall 1920, pp. xiv–xv).

Illuminating Thought Experiment

Conceptual experiments which hold constant certain endogenous variables, or vary them arbitrarily, may perform a valuable heuristic role in aiding comprehension of the attainment and character of general equilibrium, even though they are not part of the theory's logical structure. Thus, the construction of Walrasian market excess demand functions, by the mental experiment of facing each individual with the same arbitrary price vector and then aggregating, is heuristically valuable despite the fact that all market excess demands must be zero in equilibrium. In part this heuristic value comes from pertinence to the disequilibrium meta theory in which any equilibrium theory must be embedded, a meta theory which might be visualized only vaguely and informally. Mental experiments of this type have been termed 'individual' or 'ceteris paribus' experiments by Patinkin, who contrasts them with 'market' or 'mutatis mutandis' experiments in which endogenous variables are always constrained to satisfy the requirements of the underlying general equilibrium structure (Patinkin 1965, pp. 11–12).

These three different ways of invoking ceteris paribus to freeze or 'exogenize' some endogenous variables may be contrasted briefly by saying that the first views partial-equilibrium theory as sometimes preferable to general-equilibrium theory, the second regards partial-equilibrium theory as an interim step towards general-equilibrium theory, and the third interprets ceteris-paribus experiments as heuristic aids sustaining general-equilibrium theory.

The partial-equilibrium approach is closely associated with Marshall, who popularized its use, although Cournot (1838) among others had employed it previously. But Marshall's *methodological* discussion of the use of ceteris paribus restrictions arose in the narrower context of his time-period analysis, which is conducted within a framework already partial-equilibrium in character (Marshall 1920, pp. 366–80). Considering a single industry (his example is fishing), he imprisons in the pound of ceteris paribus those variables, exogenous or endogenous, whose movement is very rapid or very slow compared with those whose equilibrium and comparative-static properties he wishes to explore. The aim is to gain rough insight into likely time paths, given that explicit dynamic analysis is not feasible (see Viner 1953, p. 206).

The use, other than for frank approximations, of ceteris paribus assumptions which conflict with underlying general-equilibrium requirements (that is, the use of individual rather than market experiments) has been attacked as illogical or misleading by Friedman (1949) and Bailey (1954) in the context of demand functions, and by Buchanan (1958) more generally. A judicious assessment and summing up is provided by Yeager (1960).

Applications of ceteris paribus ideas to growth paths rather than stationary equilibria have been pioneered by Fisher and Ando (1962).

In closing, mention might be made of the classical notion of 'disturbing causes' as set out by J.S. Mill (1844, Essay V). Any deductive theorist

who regards his assumptions as true, rather than mere means for generating refutable statements, must view his (valid) deductions as also true in the absence of disturbing causes not allowed for in his assumptions (see Keynes 1891, pp. 204–13). Are such disturbing causes to be viewed as ruled out by a ceteris paribus assumption? According to Mill, they are in the statement of general economic theory (when, for example, other motives than the pursuit of wealth are excluded) but not in its specific applications, when due allowance must be made *ex ante* for all likely disturbing causes. Thus, the ruling out of disturbing causes is meant as nothing but a device to permit statement and development of a common theoretical skeleton which must be fleshed out whenever specific use is made of it.

See Also

► Marshall, Alfred (1842–1924)

Bibliography

Bailey, M.J. 1954. The Marshallian demand curve. *Journal of Political Economy* 62: 255–261.
Buchanan, J.M. 1958. Ceteris paribus: Some notes on methodology. *Southern Economic Journal* 24 (January): 259–270.
Cournot, A.A. 1838. *Mathematical principles of the theory of wealth*. Trans., New York: Macmillan, 1897.
Fisher, F.M., and A.K. Ando. 1962. Two theorems on ceteris paribus in the analysis of dynamic systems. *American Political Science Review* 56: 108–113.
Friedman, M. 1949. The Marshallian demand curve. *Journal of Political Economy* 57: 463–495. Repr. in M. Friedman, Essays in Positive Economics, Chicago: University of Chicago Press, 1953.
Keynes, J.N. 1891. *The scope and method of political economy*. London: Macmillan.
Marshall, A. 1920. *Principles of economics*. Vol. 1. 8th ed. London: Macmillan.
Mill, J.S. 1844. *Essays on some unsettled questions of political economy*. London: Parker.
Patinkin, D. 1965. *Money, interest and prices*. 2nd ed. New York: Harper & Row.
Viner, J. 1953. Cost curves and supply curves. In American Economic Association, readings in price theory. Homewood, IL: Irwin. First published in *Zeitschrift für Nationalökonomie* 3 (September 1931), 23–46.
Yeager, L.B. 1960. Methodenstreit over demand curves. *Journal of Political Economy* 68: 53–64.

Ceva, Giovanni (1647/48–1734)

Giancarlo Gandolfo

Keywords
Ceva, G.; Gresham's Law; Mathematical economics; Quantity theory of money

JEL Classifications
B31

Mathematician, hydraulic engineer and mathematical economist, Ceva was born in Milan in 1647 or 1648 and died in Mantua in 1734. He studied at the University of Pisa; later he obtained a post at Gonzaga's court in Mantua, where he became the chief technician and applied his mathematical skill to technical and administrative problems.

As a mathematician he is known for the theorem (1678) concerning the concurrency of the transverse lines from the vertices of a triangle, which is named after him; his work on fluvial hydraulics is summed up in *Opus hydrostaticum* (1728). His studies in economics are contained in a work of 1711, where he studied monetary problems. Here we find a statement of the quantity theory of money: *ceteris paribus*, the value of money varies inversely with its quantity and directly with the number of people. The latter assertion may seem odd, but it is not if we interpret 'number of people' as a proxy for the transaction variable in the quantity theory equation (as is implicit in Ceva's Postulate II). We also find an independent statement of Gresham's Law and a study of the problems of a plurimetallic standard.

The interest of this work, however, does not lie in its economics, where no objectively new contributions are made, but in its methodological content and message. Ceva was the first to conceive, to state lucidly and to apply unhesitatingly the idea of *systematically* employing the mathematical method in economics as an indispensable tool with which to reason rigorously, to understand difficult and otherwise obscure phenomena

and to put them in order. His analytico-deductive treatment, which proceeds by definitions, postulates, remarks, propositions, theorems and corollaries, is indeed the first example of mathematical economics as we now understand it.

Selected Works

1678. *De lineis rectis se invicem secantibus statica constructio.* Mediolani. (A static construction concerning straight lines which intersect one another. Milan.)

1711. *De re numaria quoad fieri potuit geometrice tractata.* Mantuae. (On money, treated mathematically as far as has been possible. Mantua.) Reprinted, with editor's Preface by E. Masè-Dari, as *Un precursore della econometria. Il saggio di Giovanni Ceva 'De re numaria' edito in Mantova nel 1711,* Modena: Pubblicazioni della Facoltá di Giurisprudenza, 1935. French translation, with translators' Introduction and notes by G.H. Bousquet and J. Roussier, in *Revue d'histoire économique et sociale,* 1958, No. 2, 129–69.

1728. *Opus hydrostaticum.* (A work on hydrostatics.) Mantua.

Bibliography

1971. *Dictionary of Scientific Biography,* vol. 3. New York: Scribner's Sons. (On Ceva's contribution to mathematics.)

1980. *Dizionario biografico degli italiani*, vol. 24. Rome: Istituto dell'Enciclopedia Italiana. (On Ceva's life and works.)

Nicolini, F. 1878. Un antico economista matematico. *Giornale degli Economisti* 8 (1): 11–23.

Chadwick, Edwin (1800–1890)

P. S. Atiyah

Public administrator and social reformer, Sir Edwin Chadwick was born at Longsight, near Manchester, on 24 January 1800 and died at East Sheen, Surrey, on 6 July 1890. He was trained as a lawyer and qualified for the bar in 1830. His early radicalism led him into contact with the utilitarians and the reforming political economists who drew their inspiration from Ricardo. He acted as Bentham's secretary and assistant for the last two years of his life. He was also a friend of the economist Nassau Senior, and he and Senior were largely responsible for the Report which led to the complete restructuring of the Poor Law in 1834, along lines which the economists had been urging for years. For the next twenty years Chadwick was employed in a variety of public administrative positions, becoming best known for his *Report on the Sanitary Condition of the Labouring Population* (1842) which laid the foundations for modern urbanized sewerage and public health measures throughout the country, and even the world. But he was a difficult man to deal with and was eventually pensioned off by the government in 1854. He wrote a large number of pamphlets, as well as being responsible wholly or partly for many important government reports (Finer 1952).

Chadwick's principal claim to fame lies in the way in which he applied his knowledge of, and passionate commitment to, utilitarian and economic analysis to many social problems of the first half of the 19th century, but only after a minute empirical investigation of the nature of the problems. Much of his work (such as the Poor Law Report) shows the influence of the orthodox economic analysis of the times, but in some respects Chadwick was years ahead of his time. In particular, there are signs of some grasp of the problem of externalities in connection with industrial accidents costs. Chadwick wanted to throw the costs of industrial accidents incurred in the construction of the railways onto the railway companies themselves. He was struck by the heavy social costs imposed by these accidents which were not borne by the railway companies, nor by the actual construction companies, and he argued that the solution to this problem was to internalize these costs to the railway companies themselves (Lewis 1950). But it was fifty years

before workers' compensation legislation was introduced in Britain, and over a hundred before some theoretical justification was offered in economic terms for this legislation.

References

Finer, S.E. 1952. *The life and times of Sir Edwin Chadwick*. London: Methuen.
Lewis, R.A. 1950. Edwin Chadwick and the railway labourers. *Economic History Review, 2nd Series* 3(1): 107.

Chalmers, Thomas (1780–1847)

D. P. O'Brien

Keywords
Chalmers, T; Malthus's theory of population; Overproduction; Over-saving; Poor Law

JEL Classifications
B31

Chalmers was born in Anstruther, Fife, and died in Edinburgh. Though he was strongly attracted to mathematics and physics in his youth, he is famous as a theologian and economist and as an active worker in the field of poor relief. Appointed to a parish in 1803, he later moved to Glasgow, where he began a famous and influential experiment in the administration of poor relief through dividing up the large parish of St John into small units and relying on a large number of voluntary helpers. He left Glasgow to become Professor of Moral Philosophy at St Andrews in 1823; in 1828 he became Professor of Divinity at Edinburgh and in 1843 he was centrally involved in the famous ecclesiastical divisions which produced the Free Church.

Endorsing Malthus's theory of population, he argued fervently (and repetitively) that the answer to the problem lay in moral education which would, in turn, lead to moral restraint. He opposed the Poor Law: it stimulated population, and interfered with private charity, which, his Glasgow experience had convinced him, was more effective. His work on aggregate demand and gluts – he argued that there could be both overproduction and over-saving since aggregate demand could be diminished not increased in proportion to both production and saving – is generally regarded as following the work of Malthus; but the essence of the argument, in terms of his aggregate demand and employment-creating analysis of trade, is present in his 1808 pamphlet, and thus precedes Malthus's own concern with aggregate demand.

Selected Works

1808. *An inquiry into the extent and stability of national resources.* Edinburgh: Oliphant & Brown.
1821–26. *The christian and civic economy of large towns.* 3 vols. Glasgow: Chalmers & Collins.
1832. *On political economy, in connexion with the moral state and moral prospects of society.* Glasgow: Collins.

Bibliography

Blaikie, W.G. 1887. Chalmers, Thomas. In *Dictionary of national biography,* vol. 3. Oxford: Oxford University Press, 1973.
Bonar, J. 1894. Chalmers, Thomas. In *Palgrave's dictionary of political economy,* ed. H. Higgs. London: Macmillan, 1925.

Chamberlin, Edward Hastings (1899–1967)

Robert E. Kuenne

Keywords
Chamberlin, E. H.; Collusion; Excess capacity; Exploitation; Firm; Theory of; Harrod, R. F.; Imperfect competition; Industrial organization;

Large-group case; Monopolistic competition; Monopoly; Oligopoly; Product differentiation; Robinson, J. V.; Schlesinger, K.; Specialization

JEL Classifications
B31

A major innovator in modern microeconomic theory, Chamberlin was born in La Conner, Washington, on 18 May 1899, and died in Cambridge, Massachusetts, on 16 July 1967. He received his Ph.D. from Harvard in 1927, became a full professor there in 1937, and occupied the David A. Wells chair from 1951 until his retirement in 1966. He edited the *Quarterly Journal of Economics* from 1948 to 1958.

Chamberlin's career exhibits a unity of professional purpose and thematic dedication over its more than 40-year length that is rare for modern theorists. Beginning with the start of his thesis research in 1925, its publication in 1933 as the seminal *Theory of Monopolistic Competition*, and continuing through eight editions, Chamberlin devoted his life to his vision of realistic market structures as mixtures of monopoly and competition.

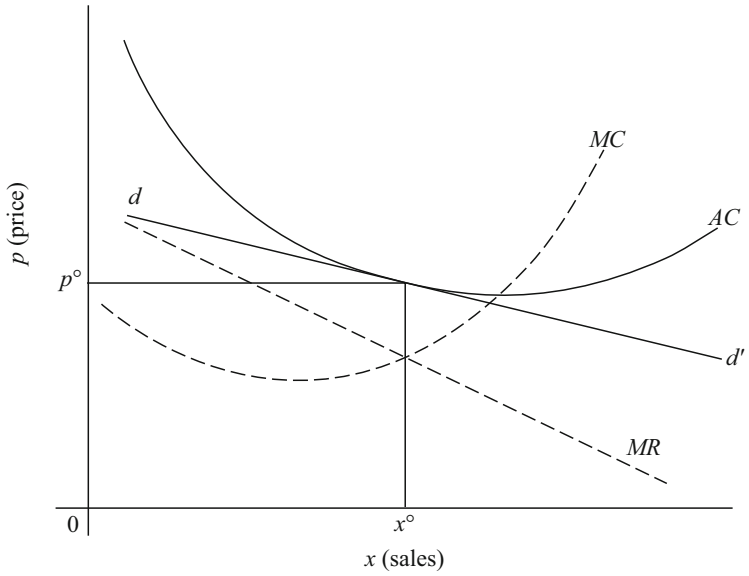
He opposed the alternative polar frameworks of pure competition and monopoly of the 1920s as unrealistic; proselytized for his merger of them at the level of the firm in both broad and narrow contexts; strove tirelessly (and rather stridently) to distinguished his concepts from Joan Robinson's similar constructs; and manned the academic ramparts in full echelon against all who sought either to criticize the concepts or, alternatively, take credit for their genesis.

In so doing, Chamberlin's broad contributions to microeconomic analysis were of fundamental and insufficiently acknowledged importance. His 'large group case' and revival of interest in oligopoly theory created the notion of market structure as a continuum between pure competition and monopoly with location dictated by numbers of firms and product differentiation. With his work he fathered modern industrial organization analysis by giving a theoretical core to what was previously institutional and anecdotal. He reoriented the interest of microeconomics from the industry to the firm, revealing the latter's target variables to include selling cost and product variation as well as price. And his frameworks led economists to comprehend the importance of differentiated oligopoly in developed economies through his emphasis upon product differentiation, his formalization of monopoly power as control over price, and his perception of the core feature of oligopolistic market structure as perceived mutual interdependence of decision making.

Monopolistic Competition Theory

In its generic sense, which Chamberlin stressed increasingly in his later career, monopolistically competitive market structures are those in which the firm feels the external compulsions of competitive forces tempered in varying degrees by a monopolistic power to price its product. Central to monopolistic competition in this wider sense is *product differentiation*, or the ability of the firm to distinguish its product in the preferences of consumers, where product is defined to include a complex of qualities in addition to those inherent in the physical good (for example, location, repair services, ambience and so on). The existence of differentiation (*a*) implies the possibility of *selling costs*, or costs aimed at adapting demand to the product (advertising, catalogues, discounts, and so on) as distinguished from *production costs*, or expenditures that adapt the product to demand, and (*b*) *product variation*, or the variability of the complex of qualities and attributes that characterize the firm's output in the mind of the consumer.

In his original presentation of monopolistic competition and into the 1940s, Chamberlin tended to identify it more narrowly with a specific market structure that isolated product differentiation as its distinctive component. This was the *large-group case* with the 'tangency solution' as the firm's long-run equilibrium position, as shown in Fig. 1. Each firm produces a slightly differentiated product which may be closely approximated by competing firms. Hence, a large number of close substitutes ensure that the firm's demand



Chamberlin, Edward Hastings (1899–1967), Fig. 1 The firm's optimal solution in the large-group case

curve is only slightly tilted from the pure competitor's horizontal position. If, for simplicity, all firms are assumed to have identical cost functions and to share sales equally (the *symmetry* assumption) then competition will reduce profit to zero by equating average cost and price at a tangency of the demand curve dd' and the average cost function *AC*. Where the tangency occurs marginal revenue *MR* will equal marginal cost *MC*. Hence, at price p° and sales x° each firm will be maximizing its profits at zero and neither entry into nor exit from the industry will occur: no internal or external force will exist to upset the long-run status quo.

Despite Chamberlin's later disclaimers, there is little doubt that the large-group case was featured as the novel contribution of his theory, and it became identified with monopolistic competition theory. But from the beginning, Chamberlin did identify a second species in the generic theory: monopolistic competition caused by fewness of sellers of a homogeneous product. In the preface to the first edition of the *Theory* he included oligopoly in the concept of monopolistic competition. Oligopoly – he coined the word independently but later recognized its prior usage in 1914 by Karl Schlesinger – in the pure (that is, undifferentiated product) case formed the mirror image of the large-group case, with small rather than large numbers of sellers and undifferentiated rather than differentiated products. Surprisingly, given the centrality of product differentiation in his thought, he had little to say about differentiated oligopoly as a composite of the two purer cases of monopolistic competition – as late as 1948 the sixth edition of the *Theory* devoted only five pages to informal discussion of it – although he realized increasingly in his later work the prominent position it held in realistic market structures.

Chamberlin's contributions to the theory of pure oligopoly were noted above in listing his broader impacts on the field. More narrowly, they were not great advances. He ignored formal treatment of collusion and tended to urge that tacit collusion would lead to joint profit maximization for pure oligopoly and to a price solution intermediate between joint profit maximization and the large-group case for differentiated oligopoly. In his later, more informal, treatment of oligopoly, however, he asserted a general tendency toward 'live-and-let-live' limitations on oligopolistic rivalry.

But from the 1950s on, Chamberlin moved away from the large-group case as the featured form of monopolistic competition theory and shifted emphasis to oligopoly in its differentiated form. In part this was an aspect of his continuing

desire to distance his theory from Joan Robinson's imperfect competition, in which she had independently developed the large-group case complete with tangency solution in the symmetry case. But, more importantly, the evolution of his thought reflected his increasing awareness that few market structures contained the uniform product competition implied by that solution. Rather, closer investigation of most realistic market structures with large numbers of sellers of slightly differentiated products revealed hierarchical clusters of oligopolistically competing firms. His book of essays (Chamberlin 1957) reveals clearly his attempt to prevent monopolistic competition theory from being too closely identified with the large-group case.

Another aspect of this later effort was the playing down of his pioneering use of marginal revenue and marginal cost curves. In denying P.W.S. Andrews's assertion that full cost pricing was antithetical to monopolistic competition, Chamberlin asserted that it was integral to that body of analysis from the beginning, since profit maximization was never an exclusive motivation of the firm – as it was in Robinson's imperfect competition.

Other Microeconomic Contributions

An implication of the large-group equilibrium illustrated in Fig. 1 is that firms would have long-run *excess capacity* in the sense that they would be operating at a production rate less than the rate associated with minimum average cost. This led to a dispute with Sir Roy Harrod, who seemed to believe that Chamberlin's results occurred because he was using short-run demand and cost curves in the large-group analysis. Harrod argued that businessmen would follow their long-run revenue and cost prospects and that excess capacity would not result. Chamberlin properly pointed out that his functions were long-run functions and that the long-run demand in Harrod's case did not attain the horizontality needed to eliminate excess capacity (Harrod 1952, Essays 7, 8; Chamberlin 1957, pp. 280–95; Kuenne 1967, pp. 67–70). Later, Chamberlin argued that excess capacity also occurred in an industry when entrants flooded in irrationally even when profits disappeared (whose counter-argument was probably what Harrod had in mind) (Chamberlin 1957, p. 290).

Chamberlin devoted a large portion of his writing to rationalizing the U-shaped average cost curve that was so fundamental to his market structures. Building upon the notion of the long-run average cost curve as the envelope of short-run average cost curves with fixed plants, he distinguished between using a fixed plant curve optimally in the short-run at its minimum-cost rate and producing a given rate of output optimally in the long-run by building an over-sized plant and using it at less than minimum cost capacity. Also, he denied that the rising portion of the long-run average cost curve was caused solely by management complexity or lumpy factors at higher output rates. In so doing, Chamberlin challenged the assertions of Knight (1921, pp. 98–9), Lerner (1944, pp. 165–7, 174–5), Stigler (1952, pp. 133, 202n.), and Kaldor (1934, p. 65n, 1935, p. 42) that, if all factors could be reduced to finely divisible units with (explicitly or implicitly assumed) constant efficiency, the average total cost would be constant as all product would be produced with optimal factor proportions. He argued that such factors would experience economies of scale as a function of factor-complex size owing to the ability to exploit specialization possibilities. These possibilities – $100 in capital might be concretized in ten shovels but $10,000 in capital might materialize as one back-hoe – permitted resource aggregates to become qualitatively different complexes with increased scale, rendering the notion of factor units with unchanged efficiency meaningless. The argument turns upon the semantics of constant efficiency units and the usefulness of the assumption, however, and was seen by most theorists to be non-illuminating and, as Chamberlin emphasized, tautological.

Two other contributions by Chamberlin are worthy of brief note. One was his destruction of Joan Robinson's notion of worker 'exploitation', because in non-purely competitive industries workers received marginal revenue product rather

than marginal value product. Chamberlin demonstrated conclusively that the difference between the two was not received by any other factor, including the entrepreneur, but was experienced as an external revenue constraint by the firm. The second, quite different, contribution was Chamberlin's role as a founder of modern experimental market research by his publication of the results of mock market operations with his students.

The Debate with Robinson

Chamberlin, like most microeconomic theorists of his generation, was thoroughly Marshallian in vision and methodology, and his innovations integrated neatly into the concerns of the post-Marshallian school. It was somewhat ironic, therefore, that Chamberlin found his major (and reluctant) opponent in Joan Robinson, as thoroughly Marshallian as himself. Chamberlin spent much of his professional life urging the fundamental divisions between his theory of monopolistic competition and Robinson's theory of imperfect competition.

The basis of the distinction changed fundamentally over his career. In the earlier objections, Chamberlin perceived correctly that Robinson's aim was to implement Sraffa's suggestion that microeconomic theory be rewritten in terms of a general theory of monopoly (Robinson 1933, p. v). In so doing, he urged, Robinson failed to achieve the true blending of monopoly and competition that his theory achieved. Robinson evolved the large-group case in every detail, but passed quickly over it in pressing on to her larger goal of creating a general theory of 'monopoly' in industries with more than one firm. To Chamberlin, who in this early period stressed the large-group case, her emphasis upon near-homogeneous commodities with some differentiation of sellers in the consumers' minds slighted the competition among differentiated products and resulted in an analysis of industry 'monopoly', very close to the one-firm monopoly of standard theory.

There was some truth in this, although Chamberlin was ungenerous to Robinson in interpreting her achievements, for in addition to her large-group case development she paralleled him in isolating selling costs and in defining two types of imperfect markets: (*a*) firms which were not alike in customers' preferences, and (*b*) oligopoly. But she saw the threat to the existence of the 'industry' that non-homogeneous products posed, and her overall goal needed that solid Marshallian construct. Chamberlin from the beginning was willing to abandon the concept and speak of 'product groups'.

However, as the large-group case came under criticism as incorporating too much of the purely competitive, and as oligopolistic structures received more attention in the literature, Chamberlin, as we have seen, shifted his ground and began to criticize Robinson for the opposite fault. The problem was, he now said, that imperfect competition failed to achieve the union of the competitive and the monopolistic because there was not enough monopoly content at the level of the firm. Implicitly, Robinson's large-group case was now focused upon for this fault, in comparison with his increasingly emphasized generic concepts that stressed oligopolistic elements.

The profession has ignored Chamberlin's strictures as distinctions without meaningful differences, and quite properly rewarded both theorists for their innovations. But the goals of the theorists were different, and, in most instances, Chamberlin's greater stress upon product differentiation and variation, selling cost and oligopoly proved to be more seminal in their professional impact.

Selected Works

1933. *The theory of monopolistic competition.* 6th edn., Cambridge, MA: Harvard University Press, 1948.

1954. (ed.) Monopoly and competition and their regulation. London: Macmillan.

1957. *Towards a more general theory of value.* New York: Oxford University Press.

Bibliography

Harrod, R.F. 1952. *Economic essays*. London: Macmillan.
Kaldor, N. 1934. The equilibrium of the firm. *Economic Journal* 44 (March): 60–76.
Kaldor, N. 1935. Market imperfection and excess capacity. *Economica* 2 (February): 33–50.
Knight, F.H. 1921. *Risk, uncertainty, and profit*. Boston: Houghton Mifflin.
Kuenne, R.E., ed. 1967. *Monopolistic competition theory: Studies in impact*. New York: Wiley.
Lerner, A.P. 1944. *The economics of control*. New York: Macmillan.
Robinson, J. 1933. *The economics of imperfect competition*. London: Macmillan.
Stigler, G.J. 1952. *The theory of price*. New York: Macmillan.

Champernowne, David Gawen (1912–2000)

F. A. Cowell

Keywords

Bayesian analysis; Champernowne, D. G.; Fat-tailed distributions; Income distribution; Scaling laws; Stochastic process models

JEL Classifications

B31

It was fortunate for the economics profession that the schoolboy Champernowne, a keen and able mathematician, was advised to read something in the school library to broaden his horizons: he chose Marshall's *Principles*.

David Champernowne was born on 9 July 1912 into an Oxford academic family. He was sent to school at Winchester and went from there as a scholar to King's College, Cambridge. While still an undergraduate he published his first paper (on 'normal numbers'). Early contact with Dennis Robertson confirmed his previous interest in economics, and he was advised by J.M. Keynes to abandon his thoughts of becoming an actuary and switch to the Economics Tripos by taking his Part II Mathematics in one year rather the normal two. He obtained firsts throughout in both subjects.

His academic career spanned the London School of Economics (1936–8) Oxford (1945–59), and Cambridge (1938–40 and 1959–78). During the war period he served with Lindemann as Assistant in the Prime Minister's Statistical Section (1940–1) and worked with Jewkes at the Ministry of Aircraft Production's Department of Statistics and Programming.

He proved to be a genuine pioneer both in economic theory and statistics. His King's fellowship dissertation (submitted in 1936, but published 27 years later in the *Economic Journal*) laid the foundations for the application of stochastic process models to the analysis of income distributions; this work has been of importance in recent economic research on fat-tailed distributions and scaling laws. His pre-war interest in Frank Ramsey's theory of probability led on to work at Oxford on the application of Bayesian analysis to autoregressive series (at a time when the Bayesian approach was decidedly unfashionable), and culminated in his major trilogy on *Uncertainty and Estimation* (1969). However although he is thought of today primarily as a theoretician, his flashes of technical insight were always been tempered with healthy doses of practical scepticism. This is evident in his early work with Beveridge on the regional and industrial distribution of employment and unemployment.

Champernowne acted as midwife to a number of major theoretical contributions over and above his own work. He provided an invaluable 'translation' to von Neumann's seminal paper on multi-sector growth. His role as behind-the-scenes expert at Cambridge over many theoretical issues is legendary: Joan Robinson acknowledged the assistance of his 'heavy artillery' in underpinning, and extending, her major work on capital and growth: A.C. Pigou's later writings on output and employment, Nicholas Kaldor's work on savings and economic growth models, and Dennis Robertson's *Principles* were all indebted to his intellectual influence.

He held Chairs at both Oxford and Cambridge, was director of the Oxford Institute of Statistics and was editor of the *Economic Journal*. He was elected Fellow of the British Academy in 1970.

Selected Works

1935. A mathematical note on substitution. *Economic Journal* 15: 246–258.
1936. Unemployment, basic and monetary: The classical analysis and the Keynesian. *Review of Economic Studies* 3: 201–216.
1938. The uneven distribution of unemployment in the United Kingdom. 1929–36, I. *Review of Economic Studies* 5: 93–106.
1939. The uneven distribution of unemployment in the United Kingdom. 1929–36, II. *Review of Economic Studies* 6: 111–124.
1945. A note on J. von Neumann's article on 'A Model of General Economic Equilibrium'. *Review of Economic Studies* 13: 10–18.
1948. Sampling theory applied to autoregressive sequences. *Journal of the Royal Statistical Society, Series B* 10: 204–242.
1952. The graduation of income distributions. *Econometrica* 20: 591–615.
1953. A model of income distribution. *Economic Journal* 63: 318–351.
1954. The production function and the theory of capital: A comment. *Review of Economic Studies* 21: 112–135.
1958. Capital accumulation and the maintenance of full employment. *Economic Journal* 68: 211–244.
1969. *Uncertainty and estimation in economics*. 3 vols. Edinburgh: Oliver and Boyd.
1971. The stability of Kaldor's 1957 model. *Review of Economic Studies* 38: 47–62.
1973. *The distribution of income between persons*. Cambridge: Cambridge University Press.
1974. A comparison of measures of income distribution. *Economic Journal* 84: 787–816.
1998. (With F. Cowall) *Economic inequality and income distribution*. Cambridge University Press.

Chandler, Alfred D. (1918–2007)

Walter A. Friedman

Keywords
Big business; Business history; Chandler, Alfred D.; Chandlerian; Economies of scale; Invisible hand; Management; Visible hand

JEL Classifications
B31

Alfred D. Chandler Jr. (15 September 1918–9 May 2007), a Pulitzer Prize-winning historian who pioneered the field of business history, was born in Guyencourt, Delaware, near Wilmington. He received a bachelor of arts degree from Harvard College in 1940 and served in the US Navy from 1941 to 1945. In the late 1940s, Chandler returned to school to study history, attending the University of North Carolina before going back to Harvard to complete his Ph.D. in 1952. He published a revision of his dissertation as his first book, *Henry Varnum Poor: Business Editor, Analyst, and Reformer* (1956). Poor (1812–1905), Chandler's paternal greatgrandfather, was the long-time editor of the *American Railroad Journal* and *Poor's Manual of Railroads of the United States*. Chandler's book explained how Poor, through his detailed reports on individual railroad companies and their operations, helped to invent the role of the modern business analyst and investment advisor.

From 1950 to 1963, Chandler taught at the Massachusetts Institute of Technology, and then left to join the history department at Johns Hopkins, where he remained until 1970. While at Hopkins, he edited the papers of President Dwight D. Eisenhower. From 1970 to 1989, Chandler was a professor at Harvard Business School, where he held the Isidor Straus Chair in Business History. While there, Chandler inaugurated the course entitled 'The Coming of Managerial Capitalism:

The United States'. He encouraged other historians to come to the school, including his successors in the Straus Chair, Thomas K. McCraw and Geoffrey Jones, and he sponsored research fellowships for graduate students and international scholars to travel to the business school's Baker Library. From the 1970s onward, he lived in a building within walking distance of the campus, in a large 17th-floor apartment overlooking the Charles River and filled with artwork, including paintings by his wife, the former Fay Martin.

Chandler was the sole author of six books and co-author or editor of more than 30 others. His most famous works focused on the rise of big business and the coming of a managerial class: *Strategy and Structure: Chapters in the History of Industrial Enterprise* (1962); *The Visible Hand: The Managerial Revolution in American Business* (1977); and *Scale and Scope: The Dynamics of Industrial Capitalism* (1990). As many commentators acknowledged, these books were so original in their approach and so impressive in their depth of research that they set the agenda for the entire field of business history for many years afterward.

The first of these, *Strategy and Structure*, analysed DuPont, General Motors, Sears and Standard Oil, and showed how each of these four companies came to adopt a multidivisional structure, or M-form, by the 1920s. No previous historian had provided such a rich account of how big businesses actually worked, or described how middle managers confronted the complexities of daily business life, filled as it was with committee meetings, budget decisions and forecasts. 'Only by showing these executives as they handled what appeared to them to be unique problems and issues can the process of innovation and change be meaningfully presented,' Chandler wrote in 1962. His detailed investigations were the basis for his influential argument that a company's strategy must shape its structure, not the other way around, as was often the case.

In *The Visible Hand*, Chandler sought to explain the rise of big business in the United States in the decades from 1840 to 1920, and to answer the question why large firms arose in some industries and not in others. Chandler argued that in industries whose firms were able to benefit from economies of scale and scope, the 'visible hand' of management came to replace the 'invisible hand' of the market in coordinating the production and distribution of goods. This was to become his most famous book, winning not only the Pulitzer, but also the Bancroft Prize and the Thomas Newcomen Book Award.

In *Scale and Scope*, Chandler branched out into comparative international history, comparing his story of the ascendancy of capitalism in the United States, from the late 19th to the mid-20th century, with the histories of Britain and Germany. Success in steel, chemicals, automobiles and other industries that emerged during the second industrial revolution, Chandler argued, was achieved through making a three-pronged investment: in mass-production facilities, in international marketing and distribution networks, and in proper management of resource allocation. While Chandler praised German industry, which had developed strong capacities in research engineering, banking, and the production of producer goods, he believed that Britain's tradition of 'personal capitalism' had prevented that country from making progress in developing large-scale industries.

These three books attracted their share of admirers throughout the world, and 'Chandlerian' business history quickly became the focus of conferences and academic papers. In 1973, Derek F. Channon published *Strategy and Structure of British Enterprise*; this was followed three years later by Gareth P. Dyas and Heinz T. Thanheiser's *The Emerging European Enterprise: Strategy and Structure in French and German Industry*. The First Fuji conference, held in Japan in January 1974, was devoted to the 'Strategy and Structure of Big Business'. Chandler's work was also central to curricula at business history units formed at the London School of Economics (in the late 1970s), and in the decades afterward at such places as the universities of Glasgow, Leeds and Reading in the United Kingdom, Bocconi University in Italy, and the Copenhagen Business School in Denmark.

Chandler also received his share of criticism, in part because of his narrow focus on the rise of big business and his relative neglect of the roles of

politics, finance, and culture in explaining the growth of the American economy. Some, including Philip Scranton (1997) and Charles F. Sabel and Michael Piore (Pire and Sabel 1984), argued that Chandler downplayed the contribution of small and medium-sized firms and overlooked the ways in which the supplanting of independent artisans and flexible manufacturers by middle managers created problems for the American economy. Chandler's most controversial book was *Scale and Scope*. British writers, in particular, bristled at Chandler's view that the preponderance of family-owned firms in the United Kingdom had contributed to that country's relative decline. Barry Supple, writing in the *Economic History Review* (1991), argued that Chandler's assumption that the American model should be the 'standard against which to assess the structural characteristics and achievements of the business systems of other countries has some pitfalls' (p. 512).

But Chandler was not an apologist for American industry, nor was he wholly enamoured with business success. He objected to many trends that were taking place in American management practice in the 1960s, including the conglomerate movement. Late in his career, he wrote admiringly of the triumph of Japanese industry over US competitors in the electronics industries in the final third of the 20th century. In the 1990s, Chandler became fascinated with the question of why some industries failed and others rose in their place. He completed his final two books, both touching on these themes, while in his 80s: *Inventing the Electronic Century: The Epic Story of the Consumer Electronics and Computer Industries* (2001); and *Shaping the Industrial Century: The Remarkable Story of the Modern Chemical and Pharmaceutical Industries* (2005).

While most historians have focused on these core books, Chandler's other works should not be neglected. He co-wrote (with Stephen Salsbury) *Pierre S. du Pont and the Making of the Modern Corporation* (1971), a long and rich primarysource study that recounts Pierre's role in making DuPont the largest US chemical and explosives company and General Motors the world's biggest car manufacturer. Chandler also edited, or co-edited, many volumes, including, with Franco Amatori and Takashi Hikino, *Big Business and the Wealth of Nations* (1997); and with James W. Cortada, *A Nation Transformed by Information: How Information has Shaped the United States from Colonial Times to the Present* (2000). He published 60 articles, many of which are listed in the bibliography of Thomas K. McCraw's edited collection, *The Essential Alfred Chandler* (1988). One extremely insightful article, published in 1994 and hence not mentioned in McCraw's volume, is his 72-page international comparative study, published in *Business History Review*, 'The competitive performance of U.S. industrial enterprises since the Second World War.' Chandler was also the general editor of the scholarly monograph series Harvard Studies in Business History, published by Harvard University Press.

Throughout his career, Chandler's work attracted attention because he continued to ask and answer broad and challenging questions. In the 1950s and 1960s, he analysed the workings of firms, while most economists and historians at the time found them uninteresting. In the 1970s and 1980s, he turned his attention to international business and to comparative analysis. The influence of Chandler's work extended far beyond the discipline of history. He made vital contributions to organizational sociology, global business studies, and to the field of strategic management. Among his many honours he had the distinction of being listed as an eminent scholar by the Academy of International Business.

Chandler's significance to business history has been summarized in McCraw (1988) and in Richard John's 1997 essay "Elaborations, revisions, dissents:

Alfred D. Chandler, Jr.'s, The Visible Hand after twenty years." In 2008, a year after Chandler's death, both *Business History Review* and *Enterprise & Society* published reflections by prominent scholars on his legacy.

Selected Works

1956. Henry Varnum poor: Business editor, analyst and reformer. Cambridge: Harvard University Press.

1962. Strategy and structure. Cambridge, Mass.: MIT Press.
1977. The visible hand: The managerial revolution in American Business. Cambridge: Belknap Press.
1990. Scale and scope: The dynamics of industrial capitalism. Cambridge: Belknap Press.
1994. The competitive performance of U.S. industrial enterprises since the Second World War. *Business History Review* 68, 255–98.
2001. Inventing the electronic century: The epic story of the consumer electronics and computer industry. New York: Free Press.
2005. Shaping the industrial century: The remarkable story of the evolution of the modern chemical and pharmaceutical industries. Cambridge: Harvard University Press.

Bibliography

Business History Review. 2008. Essays on Alfred D. Chandler. Vol. 82.
Enterprise and Society. 2008. Essays on Alfred D. Chandler. Vol. 9.
John, R. 1997. Elaborations, revisions, dissents: Alfred D. Chandler, Jr.'s, *The Visible Hand* after twenty years. *The Business History Review* 71: 151–200.
McCraw, T.K. 1988. *The essential Alfred Chandler: essays toward a historical theory of big business*. Boston: Harvard Business School Press.
Scranton, P. 1997. *Endless novelty: specialty production and American Industrialization, 1865–1925*. Princeton: Princeton University Press.
Supple, B. 1991. *Scale and scope*: Alfred Chandler and the dynamics of industrial capitalism. *The Economic History Review* 44: 500–514.

Changes in Tastes

M. S. McPherson

It is often analytically convenient to abstract from the phenomenon of changing tastes in explaining or evaluating economic phenomena. Alfred Marshall for example defended the assumption of given wants as a useful, if crude, starting point in developing utility theory. Since the 1930s, however, the assumption of given wants has hardened increasingly into dogma. A notable step in that direction was taken in 1932, when Lionel Robbins gave wide currency to the definition of economics as the study of the relations between ends and means, the ends taken as given (Robbins 1932, Chapter 2).

Nobody, of course, supposes that tastes for goods and services are literally biological givens (Stigler and Becker (1977) do deploy this claim, perhaps for its shock value, but it is natural to read them as saying that certain deep-lying wants, as for nourishment and self-esteem, are given and these deep-lying preferences interact with prices and incomes to explain changes in tastes for particular goods and services.)

The more common view is that it is efficient to divide the intellectual labour between economists who study the consequences of given tastes and sociologists, psychologists and others who explain formation of and changes in tastes. Yet on the whole, economists (with a few notable exceptions like Scitovsky (1976)) have shown little interest in what psychologists and others have had to say about preference formation and change. At a minimum, there would seem to be a need for interdisciplinary collaboration on problems, such as long run economic development or cross-national comparisons in consumption patterns, where both causes and consequences of tastes are likely to be important (see, e.g. Felix 1979). If the causes of the taste changes are non-economic, the economists on such teams might usefully concentrate on analysing their consequences.

Sometimes, however, the division of labour will not be so neat. First, tastes may change because of changes in economic variables, making tastes *endogenous* to the economic system. The alleged influence of advertising on tastes is a standard example; another is the claim that the extension of market production into traditional societies affects tastes for material consumption relative to communal 'leisure' activities (Galbraith 1958; Hefner 1983). When such interactions are empirically important, adequate economic models need to include them.

Second, when a consumer's tastes differ at different points in time, problems of temporal inconsistency in preferences and of intrapersonal preference conflict arise. A consumer who expects to have different preferences in the future faces a planning problem between his present and future self somewhat analogous to that involved in allocating resources between two consumers with different tastes. The challenge of extending the theory of rational intertemporal choice to cover such cases – with the consumer making plans he or she will actually carry out and will not regret – is considerable (Hammond 1976). Further problems arise when preferences change rhythmically or recurrently (Winston 1980). A consumer who is periodically assailed by an impulse to spend or to overeat may take steps in advance to control the consequences of those impulses, burning credit cards or locking the refrigerator. Thomas Schelling has given the label *egonomics* to the study of such strategic attempts to reconcile intrapersonal preference conflicts (Schelling 1978, 1980; Elster 1979, 1985).

The taste changes discussed so far may simply 'happen' to consumers, without their active participation or indeed sometimes without their knowledge. A further degree of complexity is introduced when it is recognized that consumers may have preferences regarding what their tastes should be. A consumer may, for example, prefer that she lose the taste for smoking, say, or acquire a taste for jogging. An adequate economic theory of consumer choice should include such second-order preferences or *meta-preferences* (Frankfurt 1971; Sen 1977). Such an extension is necessary, most simply, in order to explain the non-trivial expenditures that consumers in advanced societies make on deliberately changing their own tastes – for example, through music-appreciation classes, weight-control clinics, and SmokEnders. More broadly, tension between consumers' everyday expenditure patterns and their larger views about how they should live – what one might call their *values* rather than mere preferences (Hirschman 1985) – plays a role in explaining various features of human experience, including ambivalence, ideological commitment and the capacity for self-discipline and self-criticism. These features of behaviour will not always matter for economics, but sometimes they will. Savings behaviour and workplace relations (where willingness to identify with the organization's goals is an important variable) are obvious applications; Albert Hirschman (1982) has argued that the explanation of fluctuations in the 'taste' for political involvement relies heavily on the formation and revision of ideological metapreferences towards politics.

The fact that preferences can change raises issues for the normative as well as the explanatory dimensions of economics. One set of questions involves the role social policy should play in resolving questions of intrapersonal conflict among preferences. Society can provide support for self-control devices, for example by allowing people to enter contracts that bind their future behaviour (such as agreeing to be confined to an alcohol treatment facility). This would then involve overruling the person's future demands to be released. It is difficult to know from what standpoint to judge the effects of a decision either way on the person's welfare or liberty (Schelling 1984). Social policy can also undermine self-control, as state-run lotteries show.

Severe difficulties also arise if economic institutions or policies can change preferences. To evaluate such policies or institutions in terms of either the preferences ex ante or ex post (unless the two happen to agree) seems arbitrary. But to decide which set of preferences is itself 'objectively' better seems to most contemporary economists just as arbitrary, as well as threatening to liberty. (Many earlier economists, including notably Marshall and J.S. Mill, were much more willing to make and defend value judgements about desirable preferences). An appeal to meta-preferences may help – basing judgement on what preferences people themselves prefer. But of course meta-preferences may themselves depend on economic institutions and policies, so this solution doesn't go very deep.

From a normative standpoint, both kinds of problems – those of preference conflict and of endogenous preferences – suggest the need to move away from the strictly 'want-regarding' moral systems that underlie most neoclassical welfare economics. Such moral systems, labelled

'welfarist' by Sen (1979), exclude as morally irrelevant all information about a society except the degree to which individuals' preferences are satisfied. Other information enters the analysis only insofar as it affects the amount of preference satisfaction achieved. (Utilitarianism and Paretian welfare economics are the most important examples of welfarist moral systems.)

Relaxing the welfarist constraint admits several kinds of information that may help with 'preference change' problems. First, procedural issues about how preferences are formed may be recognized as morally important. Processes of preference formation or change that rely on misrepresentation or distortion of facts, or on emotional manipulation, may be morally downgraded irrespective of any judgement about the worth of the preferences that result (McPherson 1982, 1983). Second, measures of individual well-being may be constructed that depend on the *resources* available to an individual or on the *capabilities* he or she can exercise, rather than (only) on the amount of preference satisfaction attained (Sen 1982; Rawls 1982). This may partly free evaluations of states of affairs from dependence on existing preferences. Finally, a society can use its public deliberative processes to come to agreement on the objective value of promoting (say, through education) certain preferences or common values (Scanlon 1975). Such agreement need not presuppose that those same preferences would be valuable in other societies or times, and procedural protections could be applied to the means by which these values are promoted.

All the problems discussed in this essay – the dependence of preferences on institutions and policies, the presence of conflicting desires within the person, and the human capacity to evaluate one's own preferences – were well known to Plato and Aristotle, and were recognized by them as central issues for social theory. That the problems have remained central and largely unresolved for 25 hundred years no doubt makes some economists think it wise to define them out of the discipline, at whatever cost in realism and relevance. Others, however, welcome the resurgence of interest in problems of changing tastes as an opportunity to reestablish links with the other social sciences and with political philosophy.

See Also

- ▶ Advertising
- ▶ Preferences
- ▶ Wants

References

Elster, J. 1979. *Ulysses and the sirens: Studies in rationality and irrationality*. Cambridge: Cambridge University Press.

Elster, J. 1985. Weakness of will and the free-rider problem. *Economics and Philosophy* 1(2): 231–266.

Felix, D. 1979. De gustibus disputandum est: Changing consumer preferences in economic growth. *Explorations in Economic History* 16: 260–296.

Frankfurt, H. 1971. Freedom of the will and the concept of a person. *Journal of Philosophy* 68: 5–20.

Galbraith, J.K. 1958. *The affluent society*. Boston: Houghton-Mifflin.

Hammond, P. 1976. Changing tastes and coherent dynamic choice. *Review of Economic Studies* 43: 159–173.

Hefner, R.W. 1983. The problem of preference: Economics and ritual change in Highland Java. *Man* 17: 323–341.

Hirschman, A.O. 1982. *Shifting involvements: Private interest and public action*. Princeton: Princeton University Press.

Hirschman, A.O. 1985. Against parsimony: Three easy ways of complicating economic discourse. *Economics and Philosophy* 1: 7–21.

McPherson, M.S. 1982. Mill's moral theory and the problem of preference change. *Ethics* 92: 252–273.

McPherson, M.S. 1983. Want formation, morality, and some 'interpretive' aspects of economic inquiry. In *Social science as moral inquiry*, ed. M.S. McPherson. New York: Columbia University Press.

Rawls. 1982. Social unity and primary goods. In *Utilitarianism and beyond*, ed. A. Sen and B. Williams, 159–186. Cambridge: Cambridge University Press.

Robbins, L. 1932. *An essay on the nature and significance of economic science*. London: Macmillan. New York: St. Martin's Press, 1969.

Scanlon, T.M. 1975. Preference and urgency. *Journal of Philosophy* 72: 655–669.

Schelling, T.C. 1978. Egonomics, or the art of self-management. *American Economic Review* 68: 290–294.

Schelling, T.C. 1980. The intimate contest for self-command. *The Public Interest* 60: 94–118.

Schelling, T.C. 1984. Ethics, law and the exercise of self-command. In *Choice and consequence*, ed. T.C. Schelling. Cambridge: Harvard University Press.

Scitovsky, T. 1976. *The joyless economy*. New York: Oxford University Press.
Sen, A.K. 1977. Rational fools: A critique of the behavioral foundations of economic theory. *Philosophy and Public Affairs* 6: 317–344.
Sen, A.K. 1979. Utilitarianism and welfarism. *Journal of Philosophy* 76: 463–489.
Sen, A.K. 1982. On weights and measures: Informational constraints in social welfare analysis. In *Choice, welfare, and measurement*, ed. A.K. Sen, 226–263. Cambridge, MA: MIT Press.
Stigler, G.J., and G.S. Becker. 1977. De gustibus non est disputandum. *American Economic Review* 67: 76–90.
Winston, G. 1980. Addiction and backsliding: A theory of compulsive consumption. *Journal of Economic Behavior and Organization* 1(4): 295–324.

Chaotic Dynamics in Economics

Jess Benhabib

Abstract

A new literature in the 1980s studied the possibility that endogenous cycles and irregular chaotic dynamics resembling stochastic fluctuations could be generated by deterministic, equilibrium models of the economy, in particular in overlapping generations models and in models with infinitely lived representative agents. Other empirical studies attempted to identify whether various economic time series were generated by deterministic chaotic dynamics or stochastic fluctuations. While dynamic equilibrium models calibrated to standard parameter values can generate chaotic dynamics and endogenous cycles even under intertemporal arbitrage and without market frictions, definitive empirical evidence for chaos in economics has not yet been produced.

Keywords

Arbitrage; Chaos; Chaotic dynamics in economics; Endogenous cycles; Equilibrium cycles; Ergodic chaos; Intertemporal arbitrage; Overlapping generations model

JEL Classifications

D85

When a new literature in the 1980s showed that endogenous cycles and chaos can arise in equilibrium models in economics, it came as a surprise. The possibility of deterministic fluctuations, as opposed to fluctuations driven by exogenous stochastic shocks, had been noted in an earlier literature on business cycles, for example in the well-known multiplier-accelerator models, but not in equilibrium models of the economy with complete markets and no frictions (see for example Frisch 1933, or Samuelson,1939). Yet deterministic fluctuations in equilibrium models with predictable relative price changes should be ruled out by intertemporal arbitrage. Such considerations led to the rejection of regular endogenous cycles in favour of models whose fluctuations are driven by stochastic shocks.

The new literature on chaotic dynamics showed that deterministic cycles and chaos were indeed possible under complete intertemporal arbitrage and without any market frictions, both in standard models of overlapping generations and in calibrated models of infinitely lived representative agents (see for example Benhabib and Day 1980, 1982; Benhabib and Nishimura 1979; Grandmont 1985; and Boldrin and Montrucchio 1986). Of course, relative price fluctuations in such models had to be within the bounds allowed by the discount factor in order to be compatible with intertemporal arbitrage. (For an exploration of the relation between equilibrium cycles, chaos and discount rates in models with infinitely lived agents, see Benhabib and Rustichini 1990; Sorger 1992; Mitra 1996; and Nishimura and Yano 1996.) Furthermore, chaotic dynamics could exhibit not only deterministic endogenous cycles, but generate trajectories that are irregular, and that are statistically indistinguishable from stable linear stochastic AR1 processes (see Sakai and Tokumaru 1980).

We can usually describe a dynamical system in discrete time as chaotic if it can generate cycles of every periodicity, where a sequence $\{x_j\}$ is of period n if $x_j = x_{j+n}$ but $x_j \neq x_i$, for $j < i < n -$

1. In addition, this simple definition of chaos requires the existence of an uncountable number of initial x which give rise to bounded but aperiodic (not even asymptotically) sequences. For example the well-known hump-shaped function, $4x(1 - x)$, when iterated, generates such chaotic dynamics. The kind of chaotic dynamics described above is usually referred to as 'topological chaos'. If in addition we require that the set of initial conditions giving rise to aperiodic sequences are not simply uncountable but also have a positive (Lebesgue) measure, then we also have ergodic chaos. A useful sufficient condition to obtain topological chaos with a simple difference equation $x_{t+1} = f(x_t)$, with f continuous and mapping a closed interval into itself, is the existence of some x such that $f(f(f(x))) \leq x < f(x) < f(f(x))$. (See Li and Yorke 1975; for simple sufficient conditions for chaos in higher dimensions, see Diamond 1976, or Marotto 2005.) Note that this condition will be satisfied if the difference equation has a solution of period three. A particularly interesting feature of some dynamic systems that are chaotic is their sensitive dependence on initial conditions: initial conditions that are arbitrarily close can generate sequences that tend to diverge over time. Thus, small measurement errors in initial conditions may cause large forecasting errors, which may explain some of the difficulties associated with business-cycle forecasting.

The aperiodic but bounded trajectories that characterize chaos and exhibit sensitive dependence on initial conditions cannot continue to diverge for ever. They converge not to a point or a periodic cycle but to a bounded chaotic or 'strange' attractor. The dynamical system which induces the local separation and instability of the trajectories must eventually bend them back. The combination of local stretching and global folding generates the complex nature of the dynamics. Such dynamic behaviour is in fact a familiar theme in economics that highlights the self-correcting nature of the economic system. Shortages create incentives for increased supply; dire necessities give rise to inventions as the invisible hand guides the allocation of resources. An equally familiar theme is that of instability: the multiplier interacts with the accelerator, leading to explosive or implosive investment expenditures; self-fulfilling expectations give rise to bubbles and crashes. In combination, these two themes suggest a nonlinear system, somewhat unstable at the core, but effectively contained further out. The contribution of the new literature on chaotic dynamics starting in the early 1980s has been to demonstrate the compatibility of endogenous irregular fluctuations with equilibrium dynamics in economics.

For a very simple example of chaotic dynamics, consider a simple overlapping generations model where each generation lives two periods. The utility function of a generation born at t is $U(c_0(t), c_1(t + 1))$, where $c_0(t)$ is consumption when young and $c_1(t + 1)$ is consumption when old. This generation faces a budget constraint $c_1(t + 1) = w_1 + r(t)(w_0 - c_0(t))$, where w_0 is the endowment when young, w_1 is the endowment when old, and $r(t)$ is the rate of return on savings. The first order condition to the problem of maximizing utility subject to the budget constraint, on the assumption of interiority, yields $r(t) = \frac{U_1(c_0(t), c_1(t+1))}{U_2(c_0(t), c_1(t+1))}$. Here U_1 and U_2 denote the derivatives of the utility function U with respect to the first and second arguments. During each period t, market clearing requires that the sum of the endowments of the young and the old add up to the sum of their consumptions: $w_1 + w_0 = c_1(t) + c_0(0)$. Now consider the quadratic utility function $U(c, (t), c, (t + 1)) = ac_0(t) - 0.5b(c_0(t))^2 + c_1(t)$, $0 \leq c_0 \leq a/b$, and $a, b > 0$. If we substitute the first order condition into the budget constraint, and use the market clearing condition, the difference equation describing the dynamics is given by $c_1(t + 1) = ac_0(t)(1 - (b/a)c_0(t))$. Note that $c_0(t) \in (0, a/b)$ for all $c_0(0) \in (0, a/b)$, provided $a \leq 4$. This difference equation will exhibit chaotic dynamics in c_0 for $a \in [3.53, 4]$; $b = a$. For example, if $a = 3.83$, the difference equation has a three-period cycle for $c_0(t) = 0.1561$, where $c_0(t + 1) = 0.5096$ and $c_0(t + 2) = 0.9579$. In this simple example utility saturates at $c_0 = a/b$, but the chaotic trajectories and those with a period greater than one never attain b/a, since if $c_0(t) = b/a$, $c(t + i) = 0$ for all $i = 1, 2, \ldots$ Another simple example of an exponential utility function

that will generate chaotic dynamics in this simple overlapping generations model, for $a > 2.692$ and $w_1 > e^{a-1}$,is $U(c,(t),c,(t+1)) = A - e^{a+w_0-c_0(t)} + c_1(t)$. (See Benhabib and Day 1982, s. 3.4.)

Techniques to empirically distinguish between data generated by non-chaotic stochastic systems and deterministic chaotic systems have been developed by physicists and mathematicians (see for example Eckmann and Ruelle 1985). These techniques have been further refined into statistical tests for applications to economic data by Brock (1986) and Brock et al. (1996), among others. Very roughly, these methods exploit the idea that deterministic systems will generate trajectories that are of lower dimension than those generated by stochastic systems, which have more scattered trajectories. For example, if we consider a one-dimensional difference equation that generates chaotic dynamics, say $x_{t+1} = 4x_t\,(1 - x_t)$ for initial $x_0 \in (0, 1)$, plotting x_{t+1} against x_t will yield a curve. By contrast, if the dynamics were generated by a linear or nonlinear stochastic system with noise, the same plot would produce a scatter of points, which could not be captured by a 'relatively smooth', one-dimensional line. By formalizing this idea, we may attempt to distinguish data generated by deterministic chaotic systems and by non-chaotic stochastic systems, even without explicit knowledge of the underlying economic system generating the data. In general, however, such a method is hard to apply because, unlike data generated by scientific experiments, economic time series are often not long enough. If the order of underlying dynamical system generating the data is high-dimensional, say of the order of five or higher, or alternatively if we can only observe the realizations of a subset of the variables of the underlying economic model, distinguishing between stochastically and chaotically generated data becomes very difficult. The difficulty of empirically identifying chaos in high dimensional economic systems may be particularly important if chaotic dynamics is more likely to be manifested in disaggregated sectoral or industry data whose components, because of resource constraints or other scarcities, can move in ways that partially offset one another's cyclic or irregular movements. It would therefore be fair to say that at this point, while we know that standard dynamic equilibrium models with parameters calibrated to values often used in the literature may well generate chaotic dynamics, more definitive empirical evidence for chaos in economics has not yet been produced.

While it may be instructive to set the theories of endogenous economic fluctuations in opposition to the theories of fluctuations driven by stochastic shocks, in practice it is more helpful to consider endogenously oscillatory dynamics as complementary to stochastic fluctuations. In certain environments it may make little difference if endogenous mechanisms by themselves generate regular and irregular persistent fluctuations, or whether they give rise to damped oscillations that are sustained by stochastic shocks. On the other hand, if the underlying equilibrium system is subject to distortions and there is room for stabilization policy, correctly identifying the source of the fluctuations becomes much more important. (See for example Benhabib et al. 2002). Furthermore, recognizing the role of oscillatory dynamics may diminish our reliance on unrealistically large shocks to explain economic data, for example, in real business cycle theory.

See Also

▶ Economy as a Complex System

Bibliography

Benhabib, J., and R. Day. 1980. Erratic accumulation. *Economic Letters* 6: 113–117.

Benhabib, J., and R. Day. 1982. A characterization of erratic dynamics in the overlapping generations model. *Journal of Economic Dynamics and Control* 4: 37–55.

Benhabib, J., and K. Nishimura. 1979. The Hopf bifurcation and the existence and stability of closed orbits in multisector models of optimal economic growth. *Journal of Economic Theory* 21: 421–444.

Benhabib, J., and A. Rustichini. 1990. Equilibrium cycling with small discounting. *Journal of Economic Theory* 52: 423–432.

Benhabib, J., S. Schmitt-Grohe, and M. Uribe. 2002. Chaotic interest rate rules. *American Economic Review Papers and Proceedings* 92: 72–78.

Boldrin, M., and L. Montrucchio. 1986. On the indeterminacy of capital accumulation paths. *Journal of Economic Theory* 40: 24–39.
Brock, W. 1986. Distinguishing random and deterministic systems. *Journal of Economic Theory* 40: 68–195.
Brock, W., W. Dechert, B. LeBaron, and J. Scheinkman. 1996. A test for independence based upon the correlation dimension. *Econometric Reviews* 15: 197–235.
Diamond, P. 1976. Chaotic behavior of systems of difference equations. *International Journal of Systems Science* 7: 953–956.
Eckmann, J.-P., and D. Ruelle. 1985. Ergodic theory of chaos and strange attractors. *Review of Modern Physics* 57: 617–656.
Frisch, R. 1933. Propagation problems and impulse problems in dynamic economics. In *Economic essays in honor of Gustav Cassel.* London: Allen and Unwin. Reprinted in *Readings in business cycles,* ed. R. Gordon and L. Klein. Homewood: Richard D. Irwin, 1965.
Grandmont, J. 1985. On endogenous competitive business cycles. *Econometrica* 53: 995–1045.
Li, T., and J. Yorke. 1975. Period three implies chaos. *American Mathematical Monthly* 82: 985–992.
Marotto, F. 2005. On redefining a snap-back repeller. *Chaos, Solitons and Fractals* 25: 25–28.
Mitra, T. 1996. An exact discount factor restriction for period three cycles in dynamic optimization models. *Journal of Economic Theory* 69: 281–305.
Nishimura, K., and M. Yano. 1996. On the least upper bound of discount factors that are compatible with optimal period-three cycles. *Journal of Economic Theory* 69: 306–333.
Sakai, H., and H. Tokumaru. 1980. Autocorrelations of a certain chaos. *IEEE Transactions in Acoustic Speech Signal Process* 28: 588–590.
Samuelson, P. 1939. Interactions between the multiplier analysis and the principle of acceleration. *Review of Economic Statistics* 21: 75–78.
Sarkovskii, A. 1964. Coexistence of cycles of a continuous map of the line into itself. *Ukrains'kyi Matematychnyi Zhurnal* 16: 61–71.
Sorger, G. 1992. On the minimum rate of impatience for complicated optimal growth paths. *Journal of Economic Theory* 56: 160–179.

Characteristics

W. M. Gorman and G. D. Myles

If Eve had not insisted that 'an apple is an apple is an apple', Adam would probably have brought down the wrath of Jehovah himself by characterizing things, unsurprisingly, by their characteristics, thus bringing man-made order into chaos. There remains the problem: what characteristics? Here a little mathematics is useful. Suppose goods Y are sold only in bundles X, and that there are a_{ij} units of Y_j in X_i. The total quantity of it will be

$$y_j \sum_i a_{ij} x_i, \tag{1}$$

and the total value of a bundle X_i

$$p_i \sum_j a_{ij} q_j, \tag{2}$$

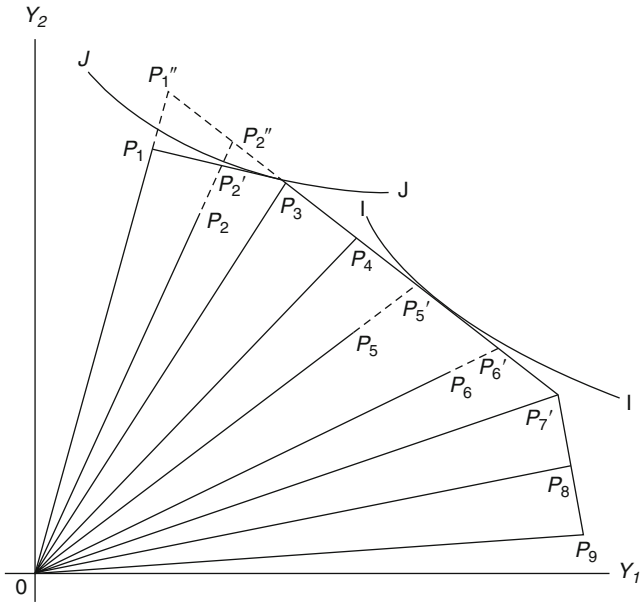
in an obvious notation, so that expenditure

$$\sum_i p_i x_i = \sum_{ij} a_{ij} q_i x_i = \sum_j q_j y_j, \tag{3}$$

for all q, x, and expenditure, m, is invariant, a fact brought to the attention of young economists at large by Samuelson when he published his *Foundations* just after the war. That immediately suggested that we might instead consider the X as goods, thought of as bundles of characteristics Y. Were we to try $y_j = f^j(x), p_i = g^i(q)$, instead of Eqs. 1 and 2, the notion that total expenditure should be invariant would yield $m = \sum g^i(q) x_i \equiv \sum q_j f^j(x)$, so that $\partial^2 m / \partial x_i \partial q_j = f^j_i(x) = g^i_j(q) = a_{ij}$ say, the subscripts denoting differentiation, and hence back to the linear characteristics model (Eqs. 1 and 2), which had already been used extensively, if implicitly, by Rowntree when studying working-class budgets in York before and after World War I, by Miss Schulz in her monthly 'human needs' budgets during World War II, by numerous nutritionists, and finally by Stigler, in a paper cited in Koopmans' Cowles Commission monograph on activity analysis in 1951 as a precursor of linear programming, a topic very fashionable among young economists at the time.

Why did demand analysts not immediately take up a model which was at once so obvious, so often used, and so in keeping with the spirit of the early Fifties? They realized, of course, that people do not eat what is good for them, so that $a_i = (a_{ij})$ each $_j$

Characteristics, Fig. 1



would have to be estimated, not taken from manuals of nutrition. Hotelling's Method of Principal Components, which had been introduced to econometricians at large by Stone in 1947, immediately suggested itself as a way of estimating them, or more precisely and relevantly, the space they span. The real problem was that the model, to be useful, would have to work with many fewer characteristics than goods to be characterized; that this seemed certain, *in practice*, to yield infinite price elasticities; and that at a time when econometric research was consistently turning up what then seemed dramatically low ones.

To see why, return to (Eq. 2). What are these mysterious q's? Nobody buys and sells raw characteristics. In fact the q's are the values which particular households put on the characteristics, and there is no reason why they should be the same for households with notably different tastes or incomes. What happens when there are two characteristics is illustrated in Fig. 1. $P_1,\ldots,P_9$ display the amounts of them available for a dollar spent on $X_1,\ldots,X_9$. One hundred times as much would be available for \$100, so that we can rescale the diagram to match the amounts particular households decide to spend on these goods as a group. Clearly nobody would buy X_2, X_5 or X_6 at these prices, so that p_5 will presumably fall to $P_5' = \left(0P_5/0P_5'\right)p_5$, for instance, at which price X_5 would be at least potentially competitive. Now consider a household whose indifference curve, appropriately scaled, is II. Its chosen bundle, y, of characteristics can be bought equally cheaply in many different ways, so that x_3, x_4, x_5, x_6 and x_7 are indeterminate. In particular it may or may not buy X_5. Should p_5 fall any further, however, it would buy a definite amount $x_5' > 0$, say, of it, while X_4 and X_6 would become uncompetitive from everybody's point of view. Hence the infinite elasticities, as we pass through *equilibrium points*.

Suppose Y_2 is a luxury. Then sufficiently rich households, like that represented by JJ, may value it sufficiently highly to *buy* X_1 and X_2, in the technical sense that it would actually buy either if its price were to fall at all. If there are too few of them to make these goods viable, their prices will have to fall for them to stay in production possibly until P_1'', P_2'', lie on P_7–P_3 extended, but possibly not. Equally, poor households may buy X_8 and X_9.

People who have never heard of a linear characteristics model often talk of the mass market, X_3–X_7, an up-market sector, X_1–X_3; and a down-market one, X_7–X_9. Even the idea of goods at the top and bottom end of the mass market, like X_3

and X_7, is quite common. That is mildly reassuring. They tend to speak of specialist markets, too, such as that for sports cars; these can easily be handled when there are more than two characteristics.

If we know the goods which fall into any one of these sectors we can fit a stochastic version of (Eq. 2) to them using, for instance, some variant of principal components analysis, given enough data. The key phrase here is 'given enough data'. Demand analysts have to deal with short inter- and autocorrelated series, subject to error, generated by the interactions between many agents, and have to be sure that the goods in question remain in the same sector throughout the period under analysis, or over the regions, for example. In practice, therefore, there have to be far fewer characteristics than goods in each sector they examine, and they commonly deal with goods bought by virtually everyone.

It is, we think, fairly generally accepted that such models are most appropriate for groups of closely related goods, to which the obvious alternative, additive separability, seems particularly badly adapted, though it has often been used by theorists interested in monopolistic competition. This is because they seem likely to differ in several ways, or, if you like, interact through several channels, which is impossible under additive separability; while the fact that they are often somewhat similar to each other suggests that linear approximations such as (Eqs. 1 and 2) may work quite well. It is consistent with the evidence from market research that households buying breakfast cereals, for instance, commmonly buy many types in quite a short period, and has been borne out by studies in the demand for related foods – although these are perhaps particularly favourable cases since we can imagine each mouthful being churned up in the stomach, the relevant components being extracted and possibly processed further to yield the characteristics in question. Labour economists quite soon began to make valiant efforts to estimate the characteristics of individual workers, or groups of workers, and to fit them to those of the functions they perform, though segmentation of the market seems more likely here, while the existence of specialization suggests that two workers with the endowment (y_1^*, y_2^*) appropriate to the job in hand may be more productive than one with $\left(y_1^* + \delta, y_2^* - \in\right)$. and another with $\left(y_1^* - \delta, y_2^* + \in\right)$.

There remains the problem of infinite elasticities.

There are two obvious routes around it: to drop linearity and hence the strong invariance property (Eq. 4), or to recognize that we are oversimplifying, so that the characteristics in question do not represent the goods perfectly. In the latter case, the natural requirement would be that the characteristics catch what these goods have in common, allowing each to have a specific value of its own, in addition, largely to be explained by the amount of it consumed. This model has been tried out with some limited success: for instance it would require at least six common characteristics Y to fit data on sources of meat protein taken from the British National Food Survey (NFS), as one plus the specific just mentioned, and either does much better than the specific alone, which is rather like an old-fashioned demand equation; the estimates were reasonably consistent in different regions; while the own price elasticities, the shadow prices of the characteristics being held constant, almost always lay between 0 and −2. Stephen Pudney estimated it by more sophisticated methods in 1980 and submitted it to formal statistical tests. It failed the tests, but did better than the alternatives he had considered – *in the particularly favourable case of food*.

If one drops linearity, the same statistical methods can be used to estimate models with

$$y_j^{\in} = \sum_i a_{ij} x_i^{\in}, \quad \text{each} \quad j, \tag{4}$$

for instance. When $\varepsilon < 1$ these are strictly concave, so that we can no longer match any particular good perfectly by baskets of others. In this case, each unit of X_i contributes towards each Y_j for which $a_{ij} \neq 0$. An alternative model

$$y_j^{\in} = \sum_i a_{ij} x_{ij}^{\in}, \quad \sum_j x_{ij} = x_i, \tag{5}$$

implies that some electricity, for instance, is used for cooking, some for lighting, some for heating,

etc. Since the consumer has to get the same value at the margin in each use, these are best characterized in the dual. In that case, they become a special, additive, example of the composite commodities described in the entry "► Separability", which were not included among the alternatives examined by Pudney. As it happens, much modern demand analysis runs in terms of composites like Eq. 5, though these are commonly interpreted as modes of behaviour appropriate to *households with particular characteristics* – of which income is the most important – rather than as *characteristics of goods*. The ε, too, are usually taken to vary from one composite commodity Y_j to another in such analyses, because poor families are commonly thought to have less flexibility in their budgets. As against this, they cannot deal with utility functions which are general in terms of the components in question, while the linear characteristics model can and does.

In practice characteristics models have commonly been applied to survey data. In the case of the NFS, for instance, each household keeps detailed records for a week. Even if this had no effect on their behaviour, they clearly buy quite different goods in different weeks quite independently of any variations in prices, and the like. The smallest units to which research workers had access contained about one hundred households: even at this level of aggregation, variability between households buying fresh beef and veal, for instance, commonly contributed less to the variance in their consumption per household than did the decision to buy it, rather than fresh mutton and lamb, and so on. This would not matter were it not for the fact that different foods may commonly be bought together: pork and apples, for instance, or strawberries and cream. Principal components analysis is based on the covariances; those between 'temporary' components may therefore contaminate the estimates of the 'permanent' characteristics, which commonly interest us most, though this should be mitigated by the use of instrumental variables as in Pudney's later work.

An alternative route would be to model the decision to buy explicitly on its own. Here the most obvious analogue is that of modal choice in transport economics, though that is considerably simplified by the fact that just one mode is chosen on any single occasion, while many different and overlapping collections of goods might be. Here the leading work was done by McFadden in 1974, building on foundations laid by Quandt and the psychologist Luce, employing an additively separable stochastic utility function to obtain choice probabilities based on the characteristics of each mode. In a manner similar to hedonic analysis these characteristics and their values for each model are assumed known, both to the consumer and to the econometrician, in contrast to the models discussed above whose aim was to identify the relevant components.

The early applications of the theory estimated, on the grounds of computational ease, multinomial logit models but this specification suffers from the 'Independence of Irrelevant Alternatives', introducing a new transport mode changes the probabilities of choosing existing modes so as to keep their ratios constant. However, alternatives are now available: Hausman and Wise discuss and estimate a conditional probit model and McFadden has introduced generalized extreme value models. At this stage, application has not yet caught up with theory.

Domenich and McFadden visualized utility being separable into seven components, one encompassing non-travel decisions and the other six travel-related decisions. These six would be made sequentially, first residential location, then vehicle ownership, trip no-trip, destination, times of day and finally mode of travel, each based on optimal decisions at lower stages. Furthermore, each choice would be made on the basis of the characteristics of the alternatives.

Despite this rich framework, studies reported in the economic journals have concentrated on either a single level or, at most, two levels of this process taking other decisions as given. This is no doubt due to the formidable data requirement of a broader analysis, detailed disaggregated data are required both on those who choose a particular mode and those who do not.

Given these difficulties, the studies that have been completed should be considered illustrative of possibilities rather than definitive. The results obtained are certainly promising, with a typical

model predicting in excess of 85 per cent of actual choice correctly. However, there is as yet no basis for evaluating the forecasting ability of the model. Winston has recently used this approach to forecast the demand for a newly established West Coast shipping line but does not provide data on the actual impact; he notes, however, that the response of incumbents to the additional competitor may invalidate his forecast.

The other main use of characteristics in empirical work is in the hedonic analysis of the prices paid for goods like houses, of which families commonly own one, so that linearity and additivity are beside the point, while the characteristics at risk are normally taken to be known. The obvious tool here is the compensating variation – that is the money $h(p, y, u) = g(p, y, u) - g(p,\overline{y}, u)$, say, required to compensate a family for living in a house with characteristics y rather than a standard house $\overline{y}$, in the obvious definition, on the demand side, and the extra cost $k(p, y) = \phi(p,\overline{y}) - \phi(p, y)$, say, of providing y rather than y, on the supply. Clearly the distribution of household characteristics is important in any equilibrium analysis, as is the structure we put on these functions. The temptation to assume separability, $h(p, y, u) = H(p, \psi(y), u)$, should probably be rejected.

How much more people are willing to pay to live in one environment rather than another is obviously important evidence in the planning of land use: how much more for a better gearbox, for directing research. Yet hedonic analysis has not provided much hard evidence so far, for lack of a secure theoretical foundation until Sherwin Rosen entered the field in the Seventies; because the potential characteristics tend to be numerous, and highly intercorrelated, while principal component analysis has often been inappropriately used; and finally because, in comparing different results, people have tended to forget what was held constant when the characteristic which interested them was varied to determine its shadow price.

A great deal of theoretical work, both in industrial and fiscal economics, has run in terms of the characteristics of the goods produced or potentially produced. Such studies are at once highly important, and difficult. Lacking hard information about the differences between goods, protagonists have tended to go for analytical convenience, or comparability with earlier work. In particular, the analogy with geographical position is often used, as has symmetry in the analysis of monopolistic competition, though there seems to be no obvious justification for them other than custom and convenience, not that either is to be despised. The results have been illustrative, therefore, of what might happen, rather than statements of what would, and in what circumstances.

There are two bridges between the empirical and theoretical schools.

Houthakker (1952) is itself purely theoretical, but clearly springs from the need to make sense of the budget data which he and Prais had been studying. He distinguishes n goods, each available in a range of qualities $\upsilon_i^- \leq \upsilon_i \leq \upsilon_i^+$, for instance at prices $p_i = a_i + b_i\upsilon_i$, where a_i and b_i are parametrically variable. He shows that a reduction in the price b_i of 'quality' accompanied by an increase in the base price a_i which leaves a family just as well off as before, leads to it buying less of i, but of a higher quality, for instance, and envisages future applications in industrial economics, in which υ_i might be a vector.

Building on earlier work with Cowling, Cubbin (1975) looked at the car market in Britain in the Sixties, combining a well-chosen characteristics model, with explicit consideration about pricing, changes in quality, and advertising; concluding from the profit margins that the industry acted oligopolistically, effective monopoly being ruled out by competition in quality.

Return to Adam. Had he known how long the first five days had lasted in the other foundation myth, and that the species around him had evolved slowly, prospering when they fitted appropriate niches, dying when they did not, he would probably have used a dynamic model, in which the characteristic space itself expanded, as new species defined new niches and new possibilities as Nalebuff and Caplin have recently argued. The economies of scale, and the political power associated with them, probably make a direct analogy with economic production inappropriate; but something like it seems to be needed when technology is changing as fast as now.

One last point. We have talked throughout as if everybody was hungry for all the characteristics. That is by no means necessarily the case. Some may be universally hated, some positively liked by some, disliked by others, though that would lead to a segmented market in the linear case.

See Also

- ▶ Demand Theory
- ▶ Goods and Commodities
- ▶ Hedonic Functions and Hedonic Indexes
- ▶ Separability

Bibliography

Caplin, P., and Nalebuff, B. 1986. Multidimensional product differentiation and price competition. *Oxford Economic Papers* 38.

Cubbin, J. 1975. Quality change and pricing behaviour in the United Kingdom car industry 1956–68. *Economica* 42: 43–58.

Domencich, T.A., and D. McFadden. 1975. *Urban travel demand*. Amsterdam: North Holland.

Gorman, W.M. 1956. A possible procedure for analysing quality differentials in the egg market. Journal paper No. 3129, Iowa Agricultural Experiment Station, and *Review of Economic Studies* (1980) 47: 843–856.

Hausman, J., and D. Wise. 1978. A conditional probit model for qualitative choice: Discrete decisions recognising interdependence and heterogeneous preferences. *Econometrica* 46: 403–426.

Hotelling, H. 1933. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* 24: 417–444. and 498–520.

Houthakker, H.S. 1952. Compensated changes in quantities and qualities consumed. *Review of Economic Studies* 19: 154–184.

Ironmonger, D.S. 1972. *New commodities and consumer behaviour*. Cambridge: Cambridge University Press.

Koopmans, T.C. (ed.). 1951. *Activity analysis of production and allocation*. New York: Wiley.

Lancaster, K.J. 1966. A new approach to consumer theory. *Journal of Political Economy* 74: 132–157.

Luce, R.D. 1959. *Individual choice behaviour*. New York: Wiley.

McFadden, D. 1981. Econometric models of probabilistic choice. In *Structural analysis of discrete data with econometric applications*, ed. C.F. Manski and D. McFadden. Cambridge, MA: MIT Press.

Quandt, R. 1956. Probabilistic theory of consumer behaviour. *Quarterly Journal of Economics* 70: 507–536.

Rowntree, P.S. 1918. *The human needs of labour*. London: Nelson.

Samuelson, P.A. 1947. *Foundations of economic analysis*. Cambridge, MA: Harvard University Press.

Schultz, T. 1943–59. Bulletin of the Oxford University Institute of Statistics, various issues.

Stigler, G.J. 1945. The cost of subsistence. *Journal of Farm Economics* 27: 303–314.

Stone, J.R.N. 1947. On the interdependence of blocks of transactions. *Journal of the Royal Statistical Society (Supplement)* 9: 1–45.

Winston, C. 1981. A multinational probit prediction of the demand for domestic ocean container services. *Journal of Transport Economics and Policy* 15: 23–42.

C

Charitable Giving

James Andreoni

Abstract

Charitable giving is a significant vehicle for providing needed goods and services around the world. In the United States, for example, charitable giving accounts for nearly two per cent of income. Moreover, the tax deduction for charitable giving is one of the oldest and most widely used tax policies in the US tax code. This article describes the known facts on charitable giving, how and why people give, and discusses the impacts of government policies on giving.

Keywords

Altruism; Charitable giving; Charitable organizations; Crowding out; Estate tax; Free rider problem; Fund-raising; Permanent income hypothesis; Philanthropy; Public goods; Self-interest; Tax deductibility; Two-stage least squares; Warm glow

JEL Classifications

H4

In 2005 charitable giving in the United States totalled over 260 billon dollars, or around 1.9 per cent of personal income, making it a significant fraction of the economy. Individual giving

accounted for 77 per cent of this total, while foundations accounted for 12 per cent, bequests for 7 per cent, and corporations for 5 per cent (Giving USA 2006). Almost 70 per cent of US households report giving to charity. While the United States typically has one of the largest and most extensively studied charitable sectors, other countries around the world also have significant philanthropy (Andreoni 2001, 2006).

There are three sets of actors in markets for charitable giving, and understanding each and their relationships to each other is essential to an understanding of charity. The first set is the donors who supply the dollars and volunteer hours to charities. The second is the charitable organizations, that is, the demand side of the market. They organize donors with fund-raising strategies, and produce the charitable goods and services with the money and time donated. The third player is the government. Governments are involved in charities in a number of ways. In many countries, including the United States, individual taxpayers may be able to deduct charitable donations from their taxable income. Governments also give directly to charities in the form of grants.

The following highlights the most important and fundamental aspects of research on charitable giving.

What Motivates Giving?

Why would a self-interested agent give away a considerable fraction of his income, often for the benefit of complete strangers? Obviously, acting unselfishly must be in his self-interest. One model of this is that the public benefits of the charity enter directly into a giver's utility function, that is, charity is a privately provided public good. This approach is advanced by Warr (1982) and Roberts (1984), who show theoretically that, if giving is a pure public good, then we would predict that government grants to charities will perfectly crowd out private donations, meaning government spending is largely ineffective. Bergstrom et al. (1986) develop this model further to provide a series of elegant derivations, including the (unrealistic) prediction that redistributions of income will be 'undone' if everyone gives to a public good. Andreoni (1988) pushes this model to its natural limits and shows that in large economies we would predict a vanishingly small fraction of people who will give to a public good, which is clearly contradicted by the statistics presented above.

For this reason, economists have felt more comfortable assuming that, in addition to caring about the total supply of charity, what could be called pure altruism, people also experience some direct private utility from the act of giving. While there are numerous models and justifications for such an assumption, they have often been gathered under the general (and slightly pejorative) term, the 'warm glow' of giving (Andreoni 1989, 1990). In large economies, in fact, it is easy to show that this motive must dominate at the margin (Ribar and Wilhelm 2002). The intuition is clear. If large numbers of others are collectively providing a substantial amount of charity, the incentive to free ride must be so overwhelming that the only remaining justification for giving is that there is some direct benefit to the act of giving.

The consequence of assuming a warm-glow motive is that we can treat individual donations as having the properties of a private good. When income is higher or when the price of giving is lower, we predict that individuals will give more.

What Is the Impact of the Tax Deduction for Charitable Giving?

Studies of the charitable deduction are aimed at understanding just how individual giving is responsive to changes in income and price. If t is the marginal tax rate faced by a giver, and if (in the United States) the person itemizes deductions, then the charitable deduction makes the effective price of a dollar of donations $1 - t$. The policy questions are how responsive is giving to the price, and is the policy successful in promoting additional giving.

Let g be the giving of the household. If the policy is effective, then the new giving received by the charity should exceed the lost revenue of the government, that is, total spending on giving

will rise with the deduction. This means $d(1-t)g/dt > 0$, which holds if $\varepsilon = [dg/d(1-t)]/[(1-t)/g] < -1$. This means that the policy is effective if giving is price-elastic, $\varepsilon < -1$. Since the first studies on giving (Feldstein and Clotfelter 1976), researchers have debated whether this 'gold standard' has been met.

Dozens of studies of this question have been undertaken. Most employ cross-sectional data, either from surveys about giving or from tax returns. Each of these data sources has advantages and weaknesses, and each presents special challenges for identification and estimation (see Triest 1998, for a careful discussion). These studies are summarized by Clotfelter (1985), Steinberg (1990), and Andreoni (2006). Prior to 1995, a consensus had formed that the income elasticity was below 1, typically in the range of 0.4 to 0.8, and that the price elasticity was below minus 1, generally in the range minus 1.1 to minus 1.3, thus meeting the gold standard. Only a few studies found giving was price-inelastic.

This consensus was upset by an important study of Randolph (1995). There are two important features of his analysis. First, he uses a panel tax returns rather than a cross section. Second, the period of his sample, 1979–89, spans two tax reforms. These reforms provide independent variation in price that can be helpful in identifying elasticities. Moreover, his instrumental variables analysis allows him to separate short-run and long-run elasticities. Contrary to the prior literature, he estimates a long-run price elasticity of only minus 0.51, meaning that the policy no longer satisfies the gold standard. Short-run elasticities, by contrast, are high, at minus 1.55. This means that givers are sophisticated at substituting giving from years of low marginal tax rates to years with high marginal tax rates. His analysis suggests that cross-sectional studies conflate short- and long-run elasticities and thus mislead policy analysts.

Auten et al. (2002) challenged Randolph's results. They use a similar (although longer) panel of tax payers, but employ a different estimation technique. Their analysis capitalizes on restrictions placed on the covariance matrices of income and price by assumptions of the permanent income hypothesis. Their analysis again returns estimates to the consensus values, with a permanent price elasticity of minus 1.26. The sensitivity of the estimates to the estimation technique and the identification strategy has left the literature unsettled as to the true values of price and income elasticities.

C

Giving by the Very Wealthy

Most of the data available, for reasons of confidentiality, exclude the very wealthy. Yet, the richest 400 US tax filers in the year 2000 accounted for about seven per cent of all individual giving in that year. Auten et al. (2000) provide a fascinating analysis of wealthy givers drawn from income tax filings at the Internal Revenue Service. Among the most interesting findings is that giving as a percentage of income rises only modestly with income, up to about four per cent for those earning over 2.5 million dollars. However, the variance in giving rises sharply. The inference is that wealthy givers are 'saving up' for larger gifts. These larger gifts may allow them to exert some control over the charity, such as providing a seat on the board of directors, or may garner a monument, such as naming a university building after the donor.

In discussing the wealthy, one must also address the effects of the estate tax on giving. Bakija et al. (2003) use 39 years' worth of federal estate tax filings to study the sensitivity of estate giving to the estate tax. They rely on variation in estate tax rates across states for identification and find that charitable giving from estates is extremely sensitive to the tax. They measure the price elasticity of estate giving to be around minus 2.0, while the 'wealth elasticity' is about 1.5. This indicates that the 2001 changes in US estate tax laws, which greatly reduce (and eventually eliminate) estate tax rates, can have huge impacts on giving.

Do Government Grants Crowd Out Individual Giving?

There are many studies on crowding out, and most show that crowding is quite small, often near zero,

and sometime even negative (Kingma 1989; Okten and Weisbrod 2000; Khanna et al. 1995; Manzoor and Straub 2005; Hungerman 2005). Payne (1998), however, noted that the government officials who approve the grants are elected by the same people who make donations to charities. Hence, positive feelings toward a charity will be represented in the preferences of both givers and the government. This positive relationship between public and private donations means that some of the prior estimates could be biased against finding crowding out.

Payne (1998) turns to two-stage least squares analysis to address this endogeneity. As an instrument for government grants she uses aggregate government transfers to individuals in the state, and finds that estimates of crowding out rise to around 50 per cent, which is significantly above the zero per cent crowing that comes when she applies prior techniques to her data. This is a significant new finding.

None of this analysis, however, has accounted for the fact that government grants may also have an impact on the fund-raising of charities. Andreoni and Payne (2003) ask what happens to a charity's fund-raising expenses when it gets a government grant. Does it fall, and by how much? They look at 14-year panel charitable organizations and find there are significant reductions in fund-raising efforts by charities after receiving government grants. This raises the possibility, therefore, that grants crowd out fund-raising, which then indirectly reduces giving, and that this may be the actual channel through which 'crowding out' occurs.

Incorporating Fund-Raising Into Research on Charitable Giving

One of the exciting new challenges for research on charitable giving is accounting for the strategic actions of charities in the analysis. This typically means understanding how charities choose fund-raising strategies, and how givers respond. A theoretical literature has emerged to provide a framework for analysing fund-raising (see Andreoni 2006, for a review). At the same time researchers have begun considering field and laboratory experiments on charitable giving. These studies look at the effectiveness of ideas proposed by the theoretical literature, and evaluate some of the standard practices of charities.

Rege and Telle (2004) and Andreoni and Petrie (2004) show in laboratory studies that the common practice of revealing the identities of givers, and reporting amounts given in categories (Harbaugh 1998), can have positive impacts on donations. Soetevent (2005) shows similar social effects in a field experiment.

List and Lucking-Reiley (2002) use a field experiment to establish that when charities require a minimum amount of contributions before a new initiate can be pursued, having a 'seed grant' can be greatly effective (Andreoni 1998), as can be guarantees of refunds in the event that the threshold of donations is not met (Bagnoli and Lipman 1989).

Landry et al. (2006), explore the use of lotteries in raising money for charities (Morgan 2000) in an actual door-to-door fundraising campaign. They find that lotteries increase giving, as expected. Perhaps surprisingly, however, they find that the physical attractiveness of the fundraiser has a significant affect on the amounts raised, and that this was at least as important as any economic incentives offered.

Conclusion

Charitable giving has been one of the perennial topics for economists. It presents challenges for the theorists to understand the motives and institutions for giving, for policy analysts to measure and identify the effects of price and income, and for experimenters to explore innovations in the market for giving. As governments become increasingly reliant on private organizations to provide public services, and as charities become increasingly sophisticated at raising money and delivering needed services, understanding the relationships among the suppliers and demanders of charity will become essential for calculating the social costs and benefits of charitable institutions.

See Also

- ▶ Altruism in Experiments
- ▶ Altruism, History of the Concept
- ▶ Crowding Out
- ▶ Externalities
- ▶ Public Finance
- ▶ Public Goods
- ▶ Tax Expenditures

Bibliography

Andreoni, J. 1988. Privately provided public-goods in a large economy the limits of altruism. *Journal of Public economics* 83: 57–73.

Andreoni, J. 1989. Giving with impure altruism: Applications to charity and Ricardian equivalence. *Journal of Political Economy* 97: 1447–1458.

Andreoni, J. 1990. Impure altruism and donations to public goods: A theory of warm-glow giving. *Economic Journal* 100: 464–477.

Andreoni, J. 1998. Toward a theory of charitable fundraising. *Journal of Political Economy* 106: 1186–1213.

Andreoni, J. 2001. The economics of philanthropy. In *International encyclopedia of social and behavioral sciences*, ed. N. Smeltser and P. Baltes. Oxford: Elsevier.

Andreoni, J. 2006. Philanthropy. In *Handbook of giving, reciprocity, and altruism*, ed. S.-C. Kolm and J. Mercier Ythier. Amsterdam: North-Holland.

Andreoni, J., and A.A. Payne. 2003. Do government grants to private charities crowd out giving or fundraising? *American Economic Review* 93: 792–812.

Andreoni, J., and R. Petrie. 2004. Public goods experiments without confidentiality: A glimpse into fundraising. *Journal of Public Economics* 88: 1605–1623.

Auten, G.E., C.T. Clotfelter, and R.L. Schmalbeck. 2000. Taxes and philanthropy among the wealthy. In *Does Atlas shrug? The economic consequences of taxing the rich*, ed. J.B. Slemrod. New York: Russell Sage.

Auten, G., H. Sieg, and C.T. Clotfelter. 2002. Charitable giving, income, and taxes: An analysis of panel data. *American Economic Review* 92: 371–382.

Bagnoli, M., and B.L. Lipman. 1989. Provision of public goods: Fully implementing the core through private contributions. *Review of Economic Studies* 56: 583–601.

Bakija, J.M., W.G. Gale, and J.B. Slemrod. 2003. Charitable bequests and taxes on inheritances and estates: Aggregate evidence from across states and time. *American Economic Review* 93: 366–370.

Bergstrom, T.C., L.E. Blume, and H.R. Varian. 1986. On the private provision of public goods. *Journal of Public Economics* 29: 25–49.

Clotfelter, C.T. 1985. *Federal tax policy and charitable giving*. Chicago: University of Chicago Press.

Feldstein, M., and C.T. Clotfelter. 1976. Tax incentives and charitable contributions in the United States: A microeconometric analysis. *Journal of Public Economics* 5: 1–26.

Giving USA. 2006. *The annual report on philanthropy for the year 2005*. New York: AAFRC Trust for Philanthropy.

Harbaugh, W.T. 1998. What do donations buy? A model of philanthropy based on prestige and warm glow. *Journal of Public Economics* 67: 269–284.

Hungerman, D.M. 2005. Are church and state substitutes? Evidence from the 1996 welfare reform. *Journal of Public Economics* 89: 2245–2267.

Khanna, J., J. Posnett, and T. Sandler. 1995. Charity donations in the UK: New evidence based on panel data. *Journal of Public Economics* 56: 257–272.

Kingma, B.R. 1989. An accurate measurement of the crowd-out effect, income effect, and price effect for charitable contributions. *Journal of Political Economy* 97: 1197–1207.

Landry, C., A. Lange, J.A. List, M.K. Price, and N.G. Rupp. 2006. Toward an understanding of the economics of charity: Evidence from a field experiment. *Quarterly Journal of Economics* 121: 747–782.

List, J.A., and D. Lucking-Reiley. 2002. The effects of seed money and refunds on charitable giving: Experimental evidence from a university capital campaign. *Journal of Political Economy* 110: 215–233.

Manzoor, S., and J. Straub. 2005. The robustness of Kingma's crowd-out estimate: Evidence from new data on contributions to public radio. *Public Choice* 123: 463–476.

Morgan, J. 2000. Financing public goods by means of lotteries. *Review of Economic Studies* 67: 761–784.

Okten, C., and B.A. Weisbrod. 2000. Determinants of donations markets. *Journal of Public Economics* 75: 255–272.

Payne, A.A. 1998. Does the government crowd-out private donations? New evidence from a sample of non-profit firms. *Journal of Public Economics* 69: 323–345.

Randolph, W.C. 1995. Dynamic income, progressive taxes, and the timing of charitable contributions. *Journal of Political Economy* 103: 709–738.

Rege, M., and K. Telle. 2004. The impact of social approval and framing on cooperation in public good situations. *Journal of Public Economics* 88: 1625–1644.

Ribar, D.C., and M.O. Wilhelm. 2002. Altruistic and joy-of-giving motivations in charitable behavior. *Journal of Political Economy* 110: 425–457.

Roberts, R.D. 1984. A positive model of private charity and public transfers. *Journal of Political Economy* 92: 136–148.

Soetevent, A.R. 2005. Anonymity in giving in a natural context: A field experiment in 30 churches. *Journal of Public Economics* 89: 2301–2323.

Steinberg, R. 1990. Taxes and giving: New findings. *Voluntas* 1: 61–79.

Triest, R.K. 1998. Econometric issues in estimating the behavioral response to taxation: A nontechnical introduction. *National Tax Journal* 51: 761–772.
Warr, P.G. 1982. Pareto optimal redistribution and private charity. *Journal of Public Economics* 19: 131–138.

Chartism

E. C. K. Gonner

The chartist movement was in its origin and its aim economic. It arose out of the economic necessities of the time, and its leaders had before them, as their ultimate object, social and industrial amelioration. To understand fully this aspect of chartism we must study the movement in its two phases: (1) from 1836 to 1839; (2) from 1840 to 1848.

1. 1836 to 1839. Three circumstances may be regarded as bringing about the chartist movement: the commercial and industrial distress immediately preceding it in time; the introduction of machinery with its effects; and the new poor law of 1834. Various men were of course variously affected by these causes; but their common action was secured by the predominant influence of one man, and the action of another, supported as he was by his colleagues. The influence referred to was that of Robert Owen, who had preached the gospel of optimism and social regeneration when all around seemed overshadowed with a gloomy present and a threatening future, and, further, urged on his followers and all with whom he came in contact, the need of education and moral elevation. It was, however, the action of Lord John Russell that brought into united action bodies so diverse in aim and constitution as the working men's association of London, the Birmingham political union, and the unions of the north, these latter being under the guidance of Feargus O'Connor. Briefly described, the first was educational and moderate, the second unstable, partly desirous of bringing about the adoption of Mr Attwood's currency scheme, and partly anxious for general industrial amelioration, while the latter formed centres for violent denunciation of the rise of machinery, and of the application of the new poor laws. All, however, hoped to attain their ends by bringing pressure to bear on parliament, itself to be rendered more amenable by a further extension of the franchise; and hence Lord John Russell's declaration against all further reform united them together and led to the formation of the national convention. The task to which this body devoted itself was mainly political, and to attain its object recourse was had first to menaces and then to open revolt. The former were disregarded and the latter was suppressed. Meantime, however, in the northern unions an almost socialistic attitude had been taken by some of the leaders. Throughout the entire movement, indeed, there had been symptoms that many were thinking of and aiming at an entire social reconstruction.
2. 1840 to 1848. The second phase of chartism differed essentially from the first. It was of smaller account in every way but one. Its strength was less, its adherents fewer, its organization less stable; but the views of its leaders were much more advanced. In theory, Bronterre O'Brien stood far ahead of any other. He was socialistic in his aims, but, unlike some of his associates, he did not confuse socialism and industrial retrogression. His schemes were, it is true, somewhat immature, but he may be described as feeling about for a new social organization. Feargus O'Connor, on the other hand, was neither so consistent nor so advanced in his aims. Thus at one time he was advocating the claims of the 'National Charter Association', for so the organization of the chartists was called, while at another, in defiance of the advice of his associates, he advocated a new scheme for bringing the people into connection with the land. In opposing the Anti-Corn Law League, it should be noticed, however, that he based his antagonism on the need which he alleged to exist of general

social reconstruction (see especially speech, 5 August 1844). But the direct effect of the agitation at this period was small. Discussions among the leaders and mutual accusations 'of interested motives' diminished their following, and it was to little or no purpose that O'Connor sought to win them back by his apparent advocacy of their interests in a periodical called *Labour*, or by his national land scheme. The latter, as a matter of fact, was financially unsound. The movement failed. That the leaders were really in earnest in their agitation is probable from the circumstances which have been alluded to, as also from their decided refusal to form any alliance with the middle class, or capitalist, reformers of Birmingham.

In its two phases, then, chartism was of economic importance. During the earlier period it aimed at economic regeneration; during the second, it not only aimed at this, but assumed a socialistic character.

Bibliography

Gammage, R.G. 1894. *History of the chartist movement, 1837–54*. Newcastle-on-Tyne.

Chartism: The Points of the Charter

G. Wallace

The Charter itself was a document in the form of an act of parliament, drafted by Francis Place from materials supplied by William Lovett. Its proposals were always summed up under six heads or 'points' viz. Universal, i.e. adult male, Suffrage, the Ballot, Annual Parliaments, Payment of Members, Equal Electoral Districts, and Abolition of Property Qualification. No one of these proposals was in any sense new, and the great majority of them had been continuously agitated for more than fifty years. The Duke of Richmond introduced a proposal for adult suffrage and equal electoral districts into the House of Lords in 1780. All or nearly all the charter 'points' were adopted by the Society of the Friends of the People, and the Corresponding Society in the earlier years of the French Revolution, and by that Edinburgh Convention for taking part in which Muir and Palmer were sentenced in 1793. The 'points' were generally spoken of as the Duke of Richmond's, or Sir Francis Burdett's, or Major Cartwright's 'plan of radical reform', and were undisguisedly intended by all their working class supporters to be used for bringing about economic as well as political equality. During the ten years following the French war every period of high prices and low wages produced a fierce agitation for 'radical reform' in the manufacturing districts and sometimes also in London. In 1830–32 the 'plan' was for a time given up in favour of the Reform Bill, but in London amendments in favour of universal suffrage were carried at the public meetings held in support of Lord Grey's bill. These were generally moved by members of the 'Rotunda Gang', or national Union of the Working Classes, many of whom had been personal disciples of Robert Owen. The reformers of 1790–1820 had advocated Tom Paine's proposal of a graduated income tax, or had been followers of 'Spence's plan' of land municipalization. These men went further, and were strongly though vaguely socialistic in tone. Place describes them as filled with bitter notions of animosity against everybody who did not concur in the absurd notions they entertained, that everything which was produced belonged to those who by their labour produced it, and ought to be shared among them; that there ought to be no accumulation of capital in the hands of any one to enable him to employ others as labourers, and thus by becoming a *master* make slaves of others under the name of workmen, to take from them the produce of their labour, to maintain themselves in idleness and luxury while their slaves were ground down to the earth or left to starve. They denounced every one who dissented from these

notions as a *political economist* under which appellation was included the notion of a bitter foe to the working classes – enemies who deserved no mercy at their hands.

Place also gives a good specimen of their teaching in a song published about this time:

> 'Wages should form the price of goods,
> Yes, wages should be all;
> Then we who work to make the goods
> Should justly have them all.
> But, if the price be made of rent,
> Tithes, taxes, profits all,
> Then we who work to make the goods
> Shall have – just none at all.'

From among these men came Lovett, Cleave, Hetherington, and others who were afterwards leaders of the chartist movement. It is significant that their organization was called successively the 'British Association for Co-operative Knowledge' (i.e. of Robert Owen's principles) in 1829; 'The Metropolitan Trades Union' in 1830, when one of their declared objects was to 'enhance the value of labour by diminishing the hours of employment', and 'The National Union of the Working Classes', for nominally political purposes, in 1831.

After the complete failure of the chartist movement in 1848, working-class reformers generally returned to the work of co-operation and trade-unionism, so that the economic side of the agitation which carred the Reform Bills of 1867 and 1884 was not so apparent as the political side. But the bill of 1867 was opposed on economic grounds by Robert Lowe (afterwards Lord Sherbrooke), Lord Shaftesbury, and others. Lord Shaftesbury on that occasion said:

I am sure that a large proportion of the working classes have a deep and solemn conviction – and I have found it among working people of religious views – that property is not distributed as property ought to be; that some checks ought to be kept on the accumulation of property in single hands; that to take away, by a legislative enactment, that which is in excess, with a view to bestow it on those who have insufficient means, is not a breach of any law, human or divine.

Chayanov, Alexander Vasil'evich (1888–1939)

Amiya K. Bagchi

Chayanov is the best-known exponent of the theory of peasant economy developed by the Organization and Production School of Russian agricultural economists. The latter were active from around 1905 down to the period of the New Economic Policy adopted by the Soviet regime.

Little is known of Chayanov's early life except that he probably came of genteel stock in European Russia. He came into early prominence and in 1913 was appointed assistant professor at the Agricultural Institute of Petrovskoe Razumovskoe (later renamed the Timiriyazev Agricultural Academy), near Moscow. In 1919, he was put in charge of the seminar on agricultural economics of the Timiriyazev Academy, later to be renamed once again as the Institute of Agricultural Economy. He directed the Institute until 1930 when, at the height of the collectivization campaign, he was dismissed. He is alleged to have died on 30 March 1939 at Alma-Ata (Smith 1976).

Chayanov was a tireless investigator into the conditions of agriculture in Russia in the era succeeding the Stolypin reforms and in the first ten years or more of the Soviet regime. He published numerous studies on cooperation, credit, peasant farming etc. in other European countries such as Italy, Belgium and Switzerland. But his main area of research centred on problems of Russian peasant production, including the question of cooperation among the peasant producers. He also took part in the organization of cooperation among the peasantry. In 1914 Chayanov proposed the organization of a central cooperative for the export of flax. Russia was then the leading exporter of flax in the world. In 1916–17 the Central Cooperative Association of

Flax Growers, of which Chayanov was a director, obtained a monopoly of flax exports, after signing an agreement with a private firm interested in the same field.

During and after the period of the Bolshevik Revolution, Chayanov became concerned with the appropriate form of agrarian reforms. He emphasized the diversity of production conditions in the different regions of Russia and of the needs of different types of products such as grain or flax. He also stressed the need to give the peasant adequate incentive to produce and market the crops needed by other sectors of the economy. By and large, he was against forcible collectivization and nationalization of land, and for voluntary cooperation among peasants who would retain control over their land. But in the period when the creation of state farms came on the political agenda, he worked out a locational plan for such farms and tried to figure out their optimum size.

The most complete bibliography of Chayanov's works available so far lists one hundred items of economic or agro-economic studies under his own name, spanning the years 1909–1930, sixteen items edited or with a preface by him (published over the years 1915–28), and twelve other works by him in the field of literature, history and arts, published over the period 1912–28 (Thorner et al. 1966, pp. 279–96). Chayanov was a cultured Russian intellectual, typical of his generation, and he wrote many of his 'non-professional pieces' under a variety of pseudonyms such as 'Botanik X', 'Moskovskii Botanik X', and 'Ivan Kremnev'. The utopia published by him under the last guise has been translated into English as *The Journey of my Brother Alexei to the Land of Peasant Utopia* (Kremnev 1920). However, it is the translation of two of his major theoretical studies into English in 1966 that kindled interest in his work among English-language readers (Thorner et al. 1966).

The centrepiece of Chayanov's theory is the concept of the family labour farm. Such a farm is supposed to employ only family labour on the family farm and on other activities such as crafts and services; on the other side, no part of this labour is hired out. Chayanov largely ignored the non-farm activities of the family labour farm. Then the equilibrium output of the farm was taken to be determined by the equation of the consumption needs of the family and the drudgery of effort (Kerblay 1966, p. xxxii).

Chayanov claimed that 90 per cent of the farms in Russia before the October Revolution were family labour farms. He used the area sown per family as the primary criterion for stratifying peasant households and claimed, on the basis of the data anlaysed by himself and by other researchers such as B.N. Knipovich, N.P. Makarov and S.N. Prokopovich, that it was the size of the family that determined the size of the area sown rather than the other way round. The direction of causality was established by means of a life cycle theory: a young family would have a high proportion of consumers to producers, and as the children grow up, the size of the family farm would increase at first to accommodate the growing needs of the family. The farm would in turn grow as more working hands are added to the family units. Then some of the young adults would move away, and settle down either on a portion of the partitioned family farm or on a new farm (Chayanov 1966, pp. 53–69).

Chayanov considered the family farm or the 'family economy' to be not only typical of pre-revolutionary or early post-revolution Russia, but to underlie a wide variety of economic systems (Chayanov 1924; this has been translated as 'On the theory of non-capitalist economic systems' in Thorner et al. 1966, pp. 1–28). In Chayanov's view, this family economy underlies not only the natural economy and 'the commodity economy' but is really at the basis of what he calls the feudal system, where the peasant household and the landlord's demesne form a symbiotic unity.

Chayanov's views on the structure of Russian rural society as well as his general view of the peasant economy as more or less a universal category have been challenged by his critics, who include the Russian Agrarian Marxists led by

L.N. Kritsman (1926) and later researchers (Harrison 1975, 1977a, b, 1979; Littlejohn 1977; Ennew et al. 1977; Chandra 1985). It has been claimed that once a multidimensional matrix of stratification is used, Russian rural society is found to have been highly stratified along class lines as Lenin had argued in his *Development of Capitalism in Russia* (1899) and later works. Chayanov's critics point out that changes in social structure cannot be explained by demographic factors alone so that the life cycle theory advanced by him does not have much of an empirical basis. Chayanov failed to take into account the fact that small peasants, rich peasants, and 'family labour farms' are held together in a web of market relationships, and that family labour farms are vulnerable to vagaries of the market as well as to natural or biological factors (see Bagchi 1982, ch. 6). Once capital is assumed to be mobile as between different farms and other sectors of the economy, the theoretical basis of an enduring family labour farm is thoroughly undermined. A peasant mode of production cannot have a theoretical validity either, because it ignores the relations of production that hold the peasantry together but also differentiate them in particular ways.

Although Chayanov's life cycle model would hold ideally only in a land-abundant economy, attempts have been made to adapt the model of the self-exploiting family farm to densely populated underdeveloped countries with widespread underemployment of labour (Georgescu-Roegen 1960) and to Polish feudalism (Kula 1962). The utopianism underlying some of the theories favouring small peasant farming under capitalist countries and explicitly spelled out by Chayanov (Kremnev 1920), has been assailed for its lack of realism and its reactionary overtones (Patnaik 1979). But there is no doubt that Chayanov raised a number of questions which, in combination with the work of Kautsky, Lenin, Mao and other Marxists, will provide a rich crop of research programmes on rural social structures wherever the peasantry formed or continue to form a large fraction of the population (Harrison 1979; Chandra 1985).

See Also

- ▶ Peasant Economy
- ▶ Peasants

Selected Works

1924. 'Zur Frage einer Theorie der nichtcapitalistischen Wirtschaftssysteme', von A. Tschayanoff. *Archiv für Sozialwissenschaft und Sozialpolitik* 51. Trans. as 'On the theory of non-capitalist economic systems', in Thorner, Kerblay and Smith (1966).

1966. In *A.V. Chayanov on the theory of the peasant economy,* ed. D. Thorner, B. Kerblay and R.E.F. Smith. Homewood: Richard D. Irwin.

1925. *Organizatsiia krest'ianskogo khoziaistva. Iz rabot Nauchno-Issledovatel'skogo Instituta s.-kh. ekonomii*. Moskva Tsentral'noe tovarichestvo kooperativnogo izd. Trans. as 'Peasant farm organization', in Thorner, Kerblay and Smith (1966).

References

Bagchi, A.K. 1982. *The political economy of underdevelopment*. Cambridge: Cambridge University Press.

Chandra, N.K. 1985. Peasantry as a single class. In *Truth unites: Essays in honour of Samar Sen*, ed. A. Mitra. Calcutta: Subarnarekha.

Ennew, J., P. Hirst, and K. Tribe. 1977. 'Peasantry' as an economic category. *Journal of Peasant Studies* 4(4): 295–322.

Georgescu-Roegen, N. 1960. Economic theory and agrarian economics. *Oxford Economic Papers*, vol. 12. Reprinted in N. Georgescu-Roegen, *Analytical economics: Issues and problems*. Cambridge, MA: Harvard University Press, 1966.

Harrison, M. 1975. Chayanov and the economics of the Russian peasantry. *Journal of Peasant Studies* 2(4): 389–417.

Harrison, M. 1977a. The problems of social mobility among Russian peasant households, 1880–1930. *Journal of Peasant Studies* 4(2): 127–161.

Harrison, M. 1977b. The peasant mode of production in the work of A.V. Chayanov. *Journal of Peasant Studies* 4(4): 323–336.

Harrison, M. 1979. Chayanov and the Marxists. *Journal of Peasant Studies* 7(1): 86–100.

Kerblay, B. 1966. A.V. Chayanov: Life, career, works. In Thorner, Kerblay and Smith (1966).

Kremnev, I. 1920. *Puteshestvie moego brata Alekseya v stranu krest'ianskoi utopii*, Moscow. Trans. as 'The journey of my brother Alexei to the land of peasant utopia' (introduced by R.E.F. Smith). *Journal of Peasant Studies* 4(1): 63–108, October 1976.
Kritsman, L.N.1926. *Klassovoe rassloenie sovetskoi derevni (po dannym volostnykh obsledovanii)*. Translated, condensed and edited by G. Littlejohn as 'Class stratification of the Soviet countryside'. *Journal of Peasant Studies* 11(2): 85–143, January 1984.
Kula, W. 1962. *Teoria ekonomiczna ustroju feudalnego*. Warsaw: Państwowe Wydawnictwo Naukowe. Translated into English from the Italian edition by L. Garner as *An Economic Theory of the Feudal System*. London: New Left Books, 1976.
Littlejohn, G. 1977. Chayanov and the theory of peasant economy. In *Sociological theories of the economy*, ed. B. Hindess. London: Macmillan.
Patnaik, U. 1979. Neo-populism and Marxism: The Chayanovian view of the agrarian question and its fundamental fallacy. *Journal of Peasant Studies* 6(4): 375–420.
Smith, R.E.F. 1976. Introduction. *Journal of Peasant Studies* 4(1): 1–8.

Cheap Money

Susan Howson

'By a long-established convention the rate of discount or the short-term rate of interest is called the "price" of money, so that "dear money" means a high rate, "cheap money" a low rate' (Hawtrey 1938, p. 28n). By the time Hawtrey was writing, however, the meaning of cheap money was changing, as a result of changes in both economic theory and monetary policy, to include low *long*-term interest rates. In the late 20th century money has not often been cheap in either sense, so that cheap or cheaper money now usually refers simply to a fall in (real) interest rates.

In the late 19th century and early 20th century, 'cheap money' meant low money market rates of interest, the rate at which commercial bills could be discounted. Since in England these rates were strongly influenced by the Bank of England's rediscount rate (Bank Rate), which was generally higher than the market rate, a 3% Bank Rate could be regarded as the upper limit of cheap money (Hawtrey 1938, p. 133). On this criterion there was cheap money for varying periods of time in all but nine years from 1844 to 1914 (Palgrave 1903, p. 98; Hawtrey 1938, Appendix I). The Bank of England, committed to maintaining the pound sterling on the gold standard with the aid of a relatively small gold reserve, varied its rate very frequently, so that these periods were of short duration, except for the spells of cheap money that followed upon the dear money of financial crises. In 1844–5, 1848–53, 1858–60, 1867–8, 1876–7, 1893–6, and 1908–9 Bank Rate was usually *below* 3% for a year of more. The most prolonged of these spells, occurring in the last years of the 'Great Depression', was permitted by a large inflow of American gold into Britain, at a time of increasing gold production, falling prices, and high unemployment (Hawtrey 1938, pp. 110–12; Sayers 1936, ch. 1; Sayers 1976, p. 51). Bank Rate stood at 2% for 2½ years, the longest period at its historical minimum before the 1930s. In the previous decade, though Bank Rate was more variable, interest rates had also been generally low. As they were to do again in the 1930s, the British government took advantage of falling long-term rates to reduce the interest paid on a large proportion of outstanding national debt: the famous 'Goschen conversion' of 1888 reduced the interest rate on 3% Consols to 2¾% until 1903 and 2½% thereafter (Clapham 1944, Vol. 2, pp. 318–21; Spinner 1973, pp. 139–503; G.J. Goschen was Chancellor of the Exchequer).

After World War I Bank Rate changes were less frequent than before 1914, partly because a high Bank Rate was now associated with high unemployment (Committee on Currency and Foreign Exchanges after the War 1918; Moggridge 1972; Sayers 1976, chs 6, 7, and 9; Howson 1975, chs 2 and 3). At the same time Bank Rate was generally higher than before the war, having been raised and kept high to curb the postwar boom in 1919–20, and again as part of the attempts to return to and stay on the gold standard at prewar parity. It was 3% only twice

in 1919–31, in 1922–3 and 1930–31, and 2½% once, for 10 weeks in mid-1931. By the time Britain left the gold standard there was a widespread desire for 'cheap money', for the sake of both the economy and the budget. Developments in monetary theory in these years (for example, Robertson 1926; Keynes 1930) implied that low *long-term* interest rates would be needed to increase investment in fixed capital and hence income and employment, rather than just low short-term rates to boost investment in working capital (inventories) as in the older views of, say, Hawtrey (1913, 1919, 1938). In 1932 the British government embarked upon a 'cheap money policy' to provide a spell of low long-term rates as well as to enable the conversion of high interest bearing government debt contracted during World War I. This also involved the establishment of an Exchange Equalization Account (EEA) to manage the exchange rate and provide sterilization of the effects of reserve changes on the monetary base. The announcement of the conversion of £2000 million 5% War Loan 1929–47 to 3½% War Loan 1952 or after was made on 30 June 1932, when Bank Rate was reduced to 2%. Apart from a short-lived rise at the outbreak of World War II, Bank Rate remained at 2% until 7 November 1951 (Nevin 1953, 1955, ch. 3; Howson 1975, ch. 4, 1980; Sayers 1976, ch. 18).

Similar, although more complex, developments in monetary theory and policy had been taking place in the USA in the interwar years (Friedman and Schwartz 1963, chs 6, 7, 8, and 9; Chandler 1971, chs 8). On both sides of the Atlantic the persistence of low interest rates and high unemployment in the 1930s induced considerable scepticism as to the efficacy of cheap money (however defined) as well as increased confidence in the monetary authorities' power to bring it about (Sayers 1951, 1957, chs 3 and 6; Keynes 1936; Wallich 1946; Morgan 1944). The decisions to maintain cheap money during and immediately after World War II reflected the scepticism, the confidence, and the desire to avoid the high borrowing costs of World War I. Monetary policy became a matter of issuing sufficient quantities of suitable debt instruments to satisfy the public's asset preferences and allowing the money supply to expand to whatever extent was necessary to maintain the fixed pattern of interest rates (Sayers 1956, chs 5 and 7; Friedman and Schwartz 1963, ch. 10; Chandler 1971, pp. 346–8). Interest rates ranged from 3/8 on Treasury bills to 2½% for long-term government bonds in the USA, and from 1% on Treasury bills to 3% for long-term government bonds in the UK. In Britain after the war, Hugh Dalton, Chancellor of the Exchequer 1945–7, also tried to go further and pursue a 'cheaper money policy', specifically to lower interest rates for government debt by ½% all the way along the yield curve. There was soon a reaction against the monetization of debt implied in these policies, and in 1947 official support of the markets for government securities was weakened in both countries, although the cheap money policies were not finally abandoned until 1951 (Paish 1947; Sayers 1957, ch. 2; Friedman and Schwartz 1963, ch. 11; Dow 1964, ch. 2 and 9; Howson 1985).

Monetary theory and practice have changed the concept of 'cheap money' again since 1951. In a more inflationary world the importance of controlling the money supply has been recognized – in the 1970s if not before – as have the inadequacies of interest rates (short or long) as an indicator of monetary conditions. When prices are rising rapidly, money can be 'cheap' even if nominal interest rates are at historically high levels. The stance of a central bank's monetary policy is now more often represented by the rate of the growth of the money supply, rather than by interest rates.

See Also

- ▶ Credit Cycle
- ▶ Dear Money
- ▶ Monetary Policy

References

Chandler, L.V. 1971. *American monetary policy 1928–1941*. New York: Harper & Row.

Clapham, J.H. 1944. *The bank of England*. Cambridge: Cambridge University Press.
Committee on currency and foreign exchanges after the war 1918. *First Interim Report*, Cd. 9182. London: HMSO.
Dow, J.C.R. 1964. *The management of the British economy 1945–60*. Cambridge: Cambridge University Press.
Friedman, M., and A.J. Schwartz. 1963. *A monetary history of the United States 1867–1960*. Princeton: Princeton University Press.
Hawtrey, R.G. 1913. *Good and bad trade*. London: Constable & Co.
Hawtrey, R.G. 1919. *Currency and credit*. London: Longmans, Green & Co.
Hawtrey, R.G. 1938. *A century of bank rate*. London: Longmans, Green & Co.
Howson, S. 1975. *Domestic monetary management in Britain 1919–38*. Cambridge: Cambridge University Press.
Howson, S. 1980. Sterling's managed float: The operations of the exchange equalisation account, 1932–39. Princeton Studies in International Finance No. 46, November.
Howson, S. 1985. The origins of cheaper money, 1945–47. Economic History Workshop, University of Toronto.
Keynes, J.M. 1930. *A treatise on money*. London: Macmillan for the Royal Economic Society. 1971.
Keynes, J.M. 1936. *The general theory of employment, interest and money*. London: Macmillan for the Royal Economic Society. 1973.
Moggridge, D.E. 1972. *British monetary policy 1924–1931*. Cambridge: Cambridge University Press.
Morgan, E.V. 1944. The future of interest rates. *Economic Journal* 54: 340–351.
Nevin, E. 1953. The origins of cheap money, 1931–32. *Economica* 20: 24–37.
Nevin, E. 1955. *The mechanism of cheap money*. Cardiff: University of Wales Press.
Paish, F.W. 1947. Cheap money policy. *Economica* 14: 167–179.
Palgrave, R.H.I. 1903. *Bank rate and the money market*. London: John Murray.
Robertson, D.H. 1926. *Banking policy and the price level*. London: P.S. King & Son.
Sayers, R.S. 1936. *Bank of England operations 1890–1914*. London: P.S. King & Son.
Sayers, R.S. 1951. The rate of interest as a weapon of economic policy. In *Oxford studies in the price mechanism*, ed. T. Wilson and P.W.S. Andrews. Oxford: Clarendon.
Sayers, R.S. 1956. *Financial policy 1939–45*. London: HMSO.
Sayers, R.S. 1957. *Central banking after Bagehot*. Oxford: Clarendon.
Sayers, R.S. 1976. *The bank of England 1891–1966*. Cambridge: Cambridge University Press.
Spinner Jr., T.J. 1973. *George Joachim Goschen*. Cambridge: Cambridge University Press.
Wallich, H.C. 1946. The changing significance of the interest rate. *American Economic Review* 36: 761–787.

Cheap Talk

Vijay Krishna and John Morgan

C

Abstract

Cheap-talk models address the question of how much information can be credibly transmitted when communication is direct and costless. When a single informed expert, who is biased, gives advice to a decision maker, only noisy information can be credibly transmitted. The more biased the expert is, the noisier the information. The decision maker can improve information transmission by: (*a*) more extensive communication, (*b*) soliciting advice from additional experts, or (*c*) writing contracts with the expert.

Keywords

Cheap talk; Communication equilibria; Delegation principle; Games with incomplete information; Incentive contracts; Revelation principle; Signalling

JEL Classifications

C7

In the context of games of incomplete information, the term 'cheap talk' refers to direct and costless communication among players. Cheap-talk models should be contrasted with more standard signalling models. In the latter, informed agents communicate private information indirectly via their choices – concerning, say, levels of education attained – and these choices are costly. Indeed, signalling is credible precisely because choices are differentially costly – for instance, high-productivity workers may distinguish themselves from low-productivity workers by acquiring levels of education that would be too costly for the latter.

The central question addressed in cheap-talk models is the following. How much information, if any, can be credibly transmitted when

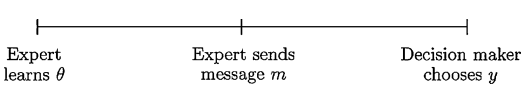
communication is direct and costless? Interest in this question stems from the fact that with cheap talk there is always a 'babbling' equilibrium in which the participants deem all communication to be meaningless – after all, it has no direct payoff consequences – and as a result no one has any incentive to communicate anything meaningful. It is then natural to ask whether there are also equilibria in which communication is meaningful and informative.

We begin by examining the question posed above in the simplest possible setting: there is a single informed party – an expert – who offers information to a single uninformed decision maker. This simple model forms the basis of much work on cheap talk and was introduced in a now classic paper by Crawford and Sobel (1982). In what follows, we first outline the main finding of this paper, namely, that while there are informative equilibria, these entail a significant loss of information. We then examine various remedies that have been proposed to solve (or at least alleviate) the 'information problem'.

The Information Problem

We begin by considering the leading case in the model of Crawford and Sobel (henceforth CS). A decision maker must choose some decision y. Her payoff depends on y and on an unknown state of the world θ, which is distributed uniformly on the unit interval. The decision maker can base her decision on the costless message m sent by an expert who knows the precise value of θ. The decision maker's payoff is $U(y, \theta) = -(y - \theta)^2$, and the expert's payoff is $V(y, \theta, b) = -(y - (\theta + b))^2$, where $b \geq 0$ is a 'bias' parameter that measures how closely aligned the preferences of the two are. Because of the tractability of the 'uniform-quadratic' specification, this paper, and indeed much of the cheap talk literature, restricts attention to this case.

The sequence of play is as follows:



What can be said about (Bayesian-perfect) equilibria of this game? As noted above, there is always an equilibrium in which no information is conveyed, even in the case where preferences are perfectly aligned (that is, $b = 0$). In such a 'babbling' equilibrium, the decision maker believes (correctly it turns out) that there is no information content in the expert's message and hence chooses her decision only on the basis of her prior information. Given this, the expert has no incentive to convey any information – he may as well send random, uninformative messages – and hence the expert indeed 'babbles'. This reasoning is independent of any of the details of the model other than the fact that the expert's message is 'cheap talk'.

Are there equilibria in which all information is conveyed? When there is any misalignment of preferences, the answer turns out to be no. Specifically,

Proposition 1 If the expert is even slightly biased, all equilibria entail some information loss.

The proposition follows from the fact that, if the expert's message always revealed the true state and the decision maker believed him, then the expert would have the incentive to exaggerate the state – in some states θ, he would report $\theta + b$.

Are there equilibria in which some but not all information is shared? Suppose that, following message m, the decision maker holds posterior beliefs given by distribution function G. The action y is chosen to maximize her payoffs given G. Because payoffs are quadratic, this amounts to choosing a y satisfying:

$$y(m) = E[\theta | m] \tag{1}$$

Suppose that the expert faces a choice between sending a message m that induces action y or an alternative message, m', that induces an action $y' > y$. Suppose further that in state θ' the expert prefers y' to y and vice versa in state $\theta < \theta'$. Since the preferences satisfy the *single-crossing* condition, $V_{y\theta} > 0$, the expert would prefer y' to y in all states higher than θ'. This implies that there is a unique state a, satisfying $\theta < a < \theta'$, in which the expert is indifferent between the two actions.

Equivalently, the distance between y and the expert's 'bliss' (ideal) action in state a is equal to the distance between action y' and the expert's bliss action in state a. Hence,

$$a + b - y = y' - (a + b) \tag{2}$$

Thus, message m is sent for all states $\theta < a$ and message m' for all states $\theta > a$.

To comprise an equilibrium where exactly two actions are induced, one would need to find values for a, y, and y' that simultaneously satisfy Eqs. (1) and (2). Since m is sent in all states $\theta < a$, from Eq. (1), $y = \frac{a}{2}$, Similarly, $y' = \frac{1+a}{2}$. Inserting these expression into eq. (2) yields

$$a = \frac{1}{2} - 2b \tag{3}$$

Equation (3) has several interesting properties. First, notice that a is uniquely determined for a given bias. Second, notice that, when the bias gets large $\left(b \geq \frac{1}{4}\right)$, there is no feasible value of a, so no information is conveyed in any equilibrium. Finally, notice that, when the expert is unbiased ($b = 0$), there exists an equilibrium where the state space is equally divided into 'high' $\left(\theta > \frac{1}{2}\right)$ and 'low' $\left(\theta < \frac{1}{2}\right)$ regions and the optimal actions respond accordingly. As the bias increases, the low region shrinks in size while the high region grows; thus, the higher the bias is, the less the information conveyed.

For all $b < \frac{1}{4}$, we constructed an equilibrium that partitions the state space into two intervals. As the bias decreases, equilibria exist that partition the state space into more than two intervals. Indeed, Crawford and Sobel (1982) showed that:

Proposition 2 All equilibria partition the state space into a finite number of intervals. The information conveyed in the most informative equilibrium is decreasing in the bias of the expert.

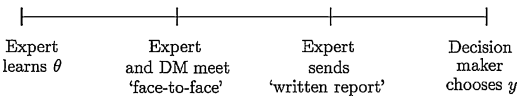
If the expert were able to commit to fully reveal what he knows, *both* parties would be better off than in any equilibrium of the game described above. With full revelation, the decision maker would choose $y = \theta$ and earn a payoff of zero, while the expert would earn a payoff of $-b^2$. It is easily verified that in any equilibrium the payoffs of both parties are lower than this. The overall message of the CS model is that, absent any commitment possibilities, cheap talk inevitably leads to information loss, which is increasing in the bias of the expert. The remainder of the article studies various 'remedies' for the information loss problem: more extensive communication, delegation, contracts, and multiple experts.

Remedies

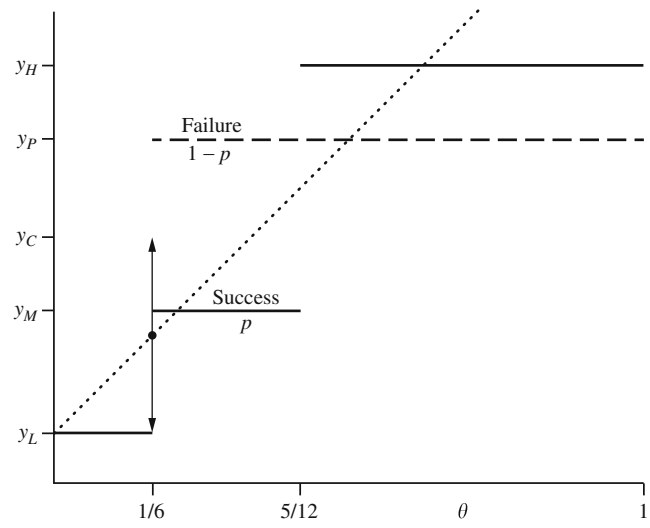
Extensive Communication

In the CS model, the form of the communication between the two parties was onesided – the expert simply offered a report to the decision maker, who then acted on it. Of course, communication can be much richer than this, and it is natural to ask whether its form affects information transmission. One might think that it would not. First, one-sided communication where the expert speaks two or more times is no better than having him speak once, since any information the expert might convey in many messages can be encoded in a single message. Now, suppose the communication is two-sided – it is a conversation – so the decision maker also speaks. Since she has no information of her own to contribute, all she can do is to send random messages, and at first glance this seems to add little. As we will show, however, random messages improve information transmission by acting as *coordinating devices*.

To see this, suppose the expert has bias $b = \frac{1}{12}$. As we previously showed, when only he speaks, the best equilibrium is where the expert reveals whether the state is above or below $\frac{1}{3}$. Suppose instead that we allow for *face-to-face* conversation – a simultaneous exchange of messages – and that the sequence of play is:



The following strategies constitute an equilibrium. The expert reveals some information at the face-to-face meeting, but there is also some randomness in what transpires. Depending on how



Cheap Talk, Fig. 1 Equilibrium with face-to-face meeting

the conversation goes, the meeting is deemed by both parties to be a 'success' or a 'failure'. After the meeting, and depending on its outcome, the expert may send an additional 'written report' to the decision maker.

During the meeting, the expert reveals whether θ is above or below $\frac{1}{6}$; he also sends some additional messages that affect the success or failure of the meeting. If he reveals that $\theta \leq \frac{1}{6}$, the meeting is adjourned, no more communication takes place, and the decision maker chooses a low action $y_L = \frac{1}{12}$ that is optimal given the information that $\theta \leq \frac{1}{6}$.

If, however, he reveals that $\theta > \frac{1}{6}$, then the written report depends on whether the meeting was a success or a failure. If the meeting is a failure, no more communication takes place, and the decision maker chooses the 'pooling' action $y_P = \frac{7}{12}$ that is optimal given that $\theta > \frac{1}{6}$. If the meeting is a success, however, the written report further divides the interval $\left[\frac{1}{6}, 1\right]$ into $\left[\frac{1}{6}, \frac{5}{12}\right]$ and $\left[\frac{5}{12}, 1\right]$. In the first subinterval, the medium action $y_M = \frac{7}{24}$ is taken and in the second sub-interval the high action $y_H = \frac{17}{24}$ is taken. The actions taken in different states are depicted in Fig. 1. The dotted line depicts the actions, $\theta + \frac{1}{12}$, that are 'ideal' for the expert.

Notice that in state $\frac{1}{6}$, the expert prefers y_L to y_P (y_L is closer to the dotted line than is y_P) and prefers y_M to y_L. Thus, if there were no uncertainty about the outcome of the meeting – for instance, if all meetings were 'successes' – then the expert would not be willing to reveal whether the state is above or below $\frac{1}{6}$; for states $\theta = \frac{1}{6} - \varepsilon$, the expert would say $\theta \in \left[\frac{1}{6}, \frac{5}{12}\right]$, thereby inducing y_M instead of y_L. If all meetings were failures, then for states $\theta = \frac{1}{6} + \varepsilon$, the expert would say $\theta < \frac{1}{6}$, thereby inducing y_L instead of y_P.

There exists a probability $p = \frac{16}{21}$ such that when $\theta = \frac{1}{6}$ the expert is indifferent between y_L and a $(p, 1-p)$ lottery between y_M and y_P (whose certainty equivalent is labelled y_C in the figure). Also, when $\theta < \frac{1}{6}$, the expert prefers y_L to a $(p, 1-p)$ lottery between y_M and y_P, and when $\theta > \frac{1}{6}$, the expert prefers a $(p, 1-p)$ lottery between y_M and y_P to y_L.

It remains to specify a conversation such that the meeting is successful with probability $p = \frac{16}{21}$. Suppose the expert sends a message (*Low*, A_i) or (*High*, A_i) and the decision maker sends a message A_j, where $i, j \in \{1, 2, \ldots, 21\}$. These messages are interpreted as follows. *Low* signals that $\theta \leq \frac{1}{6}$ and *High* signals that $\theta > \frac{1}{6}$. The A_i and A_j messages play the role of a coordinating device and

determine whether the meeting is successful. The expert chooses A_i at random and each A_i is equally likely. Similarly, the decision maker chooses A_j at random. Given these choices, the meeting is a

$$\begin{array}{ll} Success & \text{if} 0 \leq i - j < 16 \text{ or} \\ & j - i > 5 Failure \text{otherwise} \end{array}$$

For example, if the messages of the expert and the decision maker are (*High,* A_{17}) and A_5, respectively, then it is inferred that $\theta > \frac{1}{6}$ and, since $i - j = 12 < 16$, the meeting is a success. Observe that with these strategies, given any A_i or A_j, the probability that the meeting is a success is exactly $\frac{16}{21}$.

The equilibrium constructed above conveys more information than any equilibria of the CS game. The remarkable fact about the equilibrium is that this improvement in information transmission is achieved by adding a stage in which the *uninformed* decision maker also participates. While the analysis above concerns itself with the case where $b = \frac{1}{12}$, informational improvement through a 'conversation' is a general phenomenon (Krishna and Morgan 2004a):

Proposition 3 Multiple stages of communication together with active participation by the decision maker always improve information transmission.

What happens if the two parties converse more than once? Does every additional stage of communication lead to more information transmission? In a closely related setting, Aumann and Hart (2003) obtain a precise but abstract characterization of the set of equilibrium payoffs that emerge in sender–receiver games with a finite number of states and actions when the number of stages of communication is infinite. Because the CS model has a continuum of states and actions, their characterization does not directly apply. Nevertheless, it can be shown that, even with an unlimited conversation, full revelation is impossible. A full characterization of the set of equilibrium payoffs with multiple stages remains an open qst.

Delegation

A key tenet of organizational theory is the 'delegation principle', which says that the power to make decisions should reside in the hands of those with the relevant information (Milgrom and Roberts 1992). Thus, one approach to solving the information problem is simply to delegate the decision to the expert. However, the expert's bias will distort the chosen action from the decision maker's perspective. Delegation this leads to a trade-off between an optimal decision by an uninformed party and a biased decision by an informed party.

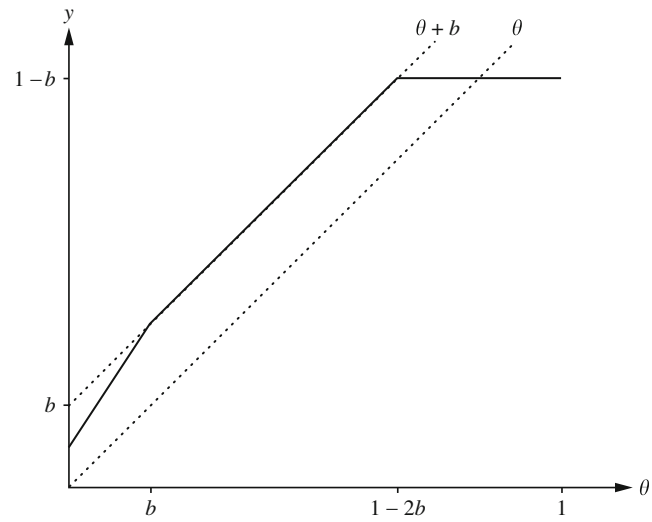
Is delegation worthwhile? Consider again an expert with bias $b = \frac{1}{12}$. The decision maker's payoff from the most informative partition equilibrium is $-\frac{1}{36}$. Under delegation, the action chosen is $y = \theta + b$ and the payoff is $-b^2 = -\frac{1}{144}$. Thus delegation is preferred. Dessein (2002) shows that this is always true:

Proposition 4 If the expert's bias is not too large $\left(b \leq \frac{1}{4}\right)$, delegation is better than all equilibria of the CS model.

In fact, by exerting only slightly more control, the decision maker can do even better. As first pointed out by Holmström (1984), the optimal delegation scheme involves limiting the scope of actions from which the expert can choose. Under the uniform-quadratic specification, the decision maker should optimally limit the expert's choice of actions to $y \in [0, 1 - b]$. When $b = \frac{1}{12}$, limiting actions in this way raises the decision maker's payoff from $-\frac{1}{144}$ to $-\frac{1}{162}$.

Optimal delegation still leads to information loss. When the expert's choice is 'capped', in high states the action is unresponsive to the state.

An application of the delegation principle arises in the US House of Representatives. Typically a specialized committee – analogous to an informed expert – sends a bill to the floor of the House – the decision maker. How it may then be amended depends on the legislative rule under effect. Under the so-called *closed rule* the floor is limited in its ability to amend the bill, while under the *open rule* the floor may freely amend the bill. Thus, operating under a closed rule is similar to delegation, while an open rule is similar to the CS model. The proposition above suggests, and Gilligan and Krehbiel (1987, 1989) have shown, that in some circumstances the floor may benefit by adopting a closed rule.



Cheap Talk, Fig. 2 An optimal contract, $b \leq \frac{1}{3}$

Contracts

Up until now we have assumed that the decision maker did not compensate the expert for his advice. Can compensation, via an incentive contract, solve the information problem? To examine this, we amend the model to allow for compensation and use mechanism design to find the optimal contract. Suppose that the payoffs are now given by

$$\begin{aligned} U(y,\theta,t) &= -(y-\theta)^2 - tV(y,\theta,b,t) \\ &= -(y-\theta-b)^2 + t \end{aligned}$$

where $t \geq 0$ is the amount of compensation.

Using the revelation principle, we can restrict attention to a direct mechanism where both t and y depend on the state θ reported by the expert. Notice that such mechanisms directly link the expert's reports to payoffs – talk is no longer cheap.

Contracts are powerful instruments. A contract that leads to full information revelation and first-best actions is:

$$t\left(\hat{\theta}\right) = 2b\left(1-\hat{\theta}\right) \quad y\left(\hat{\theta}\right) = \hat{\theta}$$

where $\hat{\theta}$ is the state reported by the expert. Under this contract, the expert can do no better than to tell the truth, that is, to set $\hat{\theta} = \theta$, and, as a consequence, the action undertaken in this scheme is the 'bliss' action for the decision maker. Full revelation is expensive, however. When $b = \frac{1}{12}$, the decision maker's payoff from this scheme is $-\frac{1}{12}$. Notice that this is worse than the payoff of $-\frac{1}{36}$ in the best CS equilibrium, which can be obtained with no contract at all. The costs of implementing the fully revealing contract outweigh the benefits.

In general, Krishna and Morgan (2004b) show:

Proposition 5 With contracts, full revelation is always feasible but never optimal.

The proposition above shows that full revelation is never optimal. No contract at all is also not optimal – delegation is preferable. What is the structure of the optimal contract? A typical optimal contract is depicted as the dark line in Fig. 2. First, notice that, even though the decision maker could induce his bliss action for some states, it is never optimal to do so. Instead, for low states ($\theta < b$) the decision maker implements a 'compromise' action – an action that lies between θ and $\theta + b$. When $\theta > b$, the optimal contract simply consists of capped delegation.

Multiple Senders

Thus far we have focused attention on how a decision maker should consult a single expert. In

many instances, decision makers consult multiple experts – often with similar information but differing ideologies (biases). Political leaders often form cabinets of advisors with overlapping expertise. How should a cabinet be constituted? Is a balanced cabinet – one with advisors with opposing ideologies – helpful? How should the decision maker structure the 'debate' among her advisors?

To study these issues, we add a second expert having identical information to the CS model. To incorporate ideological differences, suppose the experts have differing biases. When both b_1 and b_2 are positive, the experts have *like bias* – both prefer higher actions than does the decision maker. In contrast, if $b_1 > 0$ and $b_2 < 0$, then the experts have *opposing bias* – expert 1 prefers a higher action and expert 2 a lower action than does the decision maker.

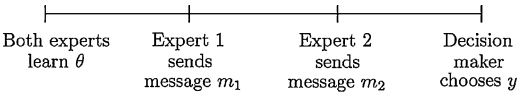
Simultaneous Talk

When both experts report to the decision maker simultaneously, the information problem is apparently solved – full revelation is now an equilibrium. To see this, suppose the experts have like bias and consider the following strategy for the decision maker: choose the action that is the more 'conservative' of the two recommendations. Precisely, if $m_1 < m_2$, choose action m_1 and vice versa if $m_2 < m_1$. Under this strategy, each expert can do no better than to report θ honestly if the other does likewise. If expert 2 reports $m_2 = \theta$, then a report $m_1 > \theta$ has no effect on the action. However, reporting $m_1 < \theta$ changes the action to $y = m_1$, but this is worse for expert 1. Thus, expert 1 is content to simply tell the truth. Opposing bias requires a more complicated construction, but the effect is the same: full revelation is an equilibrium (see Krishna and Morgan 2001b).

Notice that the above construction is fragile because truth-telling is a weakly dominated strategy. Each expert is at least as well off by reporting $m_i = \theta + b_i$ and strictly better off in some cases. Battaglini (2002) defines an equilibrium refinement for such games which, like the notion of perfect equilibrium in finite games, incorporates the usual idea that players may make mistakes. He then shows that such a refinement rules out all equilibria with full revelation regardless of the direction of the biases. While the set of equilibria satisfying the refinement is unknown, the fact that full revelation is ruled out means that simply adding a second expert does not solve the information problem satisfactorily.

Sequential Talk

Finally, we turn to the case where the experts offer advice in sequence:



Suppose that the two experts have biases $b_1 = \frac{1}{18}$ and $b_2 = \frac{1}{12}$, respectively. It is easy to verify (with the use of (2)) that, if only expert 1 were consulted, then the most informative equilibrium entails his revealing that the state is below $\frac{1}{9}$ or between $\frac{1}{9}$ and $\frac{4}{9}$, or above $\frac{4}{9}$. If only expert 2 were consulted, then the most informative equilibrium is where he reveals whether the state is below or above $\frac{1}{3}$. If the decision maker were able to consult only one of the two experts, she would be better off consulting the more loyal expert 1.

But what happens if she consults both? It turns out that, if both experts actively contribute information, then the decision maker can do no better than the following equilibrium. Expert 1 speaks first and reveals whether or not the state is above or below $\frac{11}{27}$. If expert 1 reveals that the state is above $\frac{11}{27}$, expert 2 reveals nothing further. If, however, expert 1 reveals that the state is below $\frac{11}{27}$, then expert 2 reveals further whether or not it is above or below $\frac{1}{27}$. That this is an equilibrium may be verified again by using (2) and recognizing that, in state $\frac{1}{27}$, expert 2 must be indifferent between the optimal action in the interval $\left[0, \frac{1}{27}\right]$ and the optimal action in $\left[\frac{1}{27}, \frac{11}{27}\right]$. In state $\frac{11}{27}$, expert 1 must be indifferent between the optimal action in $\left[\frac{1}{27}, \frac{11}{27}\right]$ and the optimal action in $\left[\frac{11}{27}, 1\right]$.

Sadly, by actively consulting both experts, the decision maker is worse off than if she simply ignored expert 2 and consulted only her more loyal advisor, expert 1. This result is quite general, as shown by Krishna and Morgan (2001a):

Proposition 6 When experts have like biases, actively consulting the less loyal expert never helps the decision maker.

The situation is quite different when experts have opposing biases, that is, when the cabinet is balanced. To see this, suppose that the cabinet is comprised of two equally loyal experts biases $b_1 = \frac{1}{12}$ and $b_2 = -\frac{1}{12}$. Consulting expert 1 alone leads to a partition $\left[0, \frac{1}{3}\right], \left[\frac{1}{3}, 1\right]$ while consulting expert 2 alone leads to the partition $\left[0, \frac{2}{3}\right], \left[\frac{2}{3}, 1\right]$. If instead the decision maker asked both experts for advice, the following is an equilibrium: expert 1 reveals whether θ is above or below $\frac{2}{9}$. If he reveals that the state is below $\frac{2}{9}$, the discussion ends. If, however, expert 1 indicates that the state is above $\frac{2}{9}$, expert 2 is actively consulted and reveals further whether the state is above or below $\frac{7}{9}$. Based on this, the decision maker takes the appropriate action. One may readily verify that this is an improvement over consulting either expert alone. Once again the example readily generalizes:

Proposition 7 When experts have opposing biases, actively consulting both experts always helps the decision maker.

Indeed, the decision maker can be more clever than this. One can show that, with experts of opposing bias, there exist equilibria where a portion of the state space is *fully revealed.* By allowing for a 'rebuttal' stage in the debate, there exists an equilibrium where *all* information is fully revealed.

See Also

- ▶ Agency Problems
- ▶ Signalling and Screening

Bibliography

Aumann, R., and S. Hart. 2003. Long cheap talk. *Econometrica* 71: 1619–1660.

Austen-Smith, D. 1993. Interested experts and policy advice: Multiple referrals under open rule. *Games and Economic Behavior* 5: 3–43.

Baron, D. 2000. Legislative organization with informational committees. *American Journal of Political Science* 44: 485–505.

Battaglini, M. 2002. Multiple referrals and multidimensional cheap talk. *Econometrica* 70: 1379–1401.

Crawford, V., and J. Sobel. 1982. Strategic information transmission. *Econometrica* 50: 1431–1451.

Dessein, W. 2002. Authority and communication in organizations. *Review of Economic Studies* 69: 811–838.

Gilligan, T., and K. Krehbiel. 1987. Collective decision-making and standing committees: An informational rationale for restrictive amendment procedures. *Journal of Law, Economics, and Organization* 3: 287–335.

Gilligan, T., and K. Krehbiel. 1989. Asymmetric information and legislative rules with a heterogeneous committee. *American Journal of Political Science* 33: 459–490.

Grossman, G., and E. Helpman. 2001. *Special interest politics*. Cambridge, MA: MIT Press.

Holmström, B. 1984. On the theory of delegation. In *Bayesian models in economic theory*, ed. M. Boyer and R. Kihlstrom. Amsterdam: North-Holland.

Krishna, V., and J. Morgan. 2001a. A model of expertise. *Quarterly Journal of Economics* 116: 747–775.

Krishna, V., and J. Morgan. 2001b. Asymmetric information and legislative rules: Some amendments. *American Political Science Review* 95: 435–452.

Krishna, V., and J. Morgan. 2004a. The art of conversation: Eliciting information from experts through multi-stage communication. *Journal of Economic Theory* 117: 147–179.

Krishna, V., and J. Morgan. 2004b. *Contracting for information under imperfect commitment*. Paper CPC05-051. Competition Policy Center, University of California, Berkeley.

Milgrom, P., and J. Roberts. 1992. *Economics, organization and management*. Englewood Cliffs, NJ: Prentice Hall.

Wolinsky, A. 2002. Eliciting information from multiple experts. *Games and Economic Behavior* 41: 141–160.

Chemical Industry

Ashish Arora and Alfonso Gambardella

Keywords

Chemical industry; Complementarities; Innovation; Licensing of technology; Patents; Petrochemical industry; Technical change; Vertical integration

JEL Classifications

L65

The chemical industry is among the largest manufacturing industries; its products range from acids to intermediate chemicals such as synthetic fibres and plastics, and to final products such as soaps, cosmetics, paints and fertilizers. Perhaps as a result, the chemical industry is under-studied by economists, though not by economic and business historians (e.g., Hounshell and Smith 1988).

The modern chemical industry has its origins in the discovery of synthetic dyes in Britain in the 1850s. German chemical firms such as BASF, Bayer and Hoechst soon dominated the production of synthetic dyestuffs and related organic compounds. The American chemical industry grew by exploiting the rich American natural resource endowments, initially using European technology.

After the First World War, American firms, especially Du Pont, invested in R&D. The inter-war period saw rapid product innovation in synthetic fibres, plastics, resins, adhesives, paints, and coatings, based on polymer science. To succeed commercially, these products had to be produced cheaply, which meant large-scale production and, in turn, the development of chemical engineering. The Second World War marked a watershed. The chemical industry became closely linked with the oil industry, as many chemicals used petroleum-based inputs instead of coal byproducts. The United States was the first country to develop a petrochemicals industry, mainly due to its abundant oil reserves, as well as wartime government programmes for aviation fuel and synthetic rubber.

The early advantage of the US chemical industry in petrochemicals was eroded as technologies diffused widely, first to Europe and Japan; and in the 1970s China, Taiwan and S. Korea emerged as leading producers. Increased competition, the oil shocks of the 1970s, and waning possibilities for product innovation together resulted in exit: larger, multi-product firms exited earlier, but larger plants closed later (Lieberman 1990). In addition, firms reshuffled product portfolios so as to focus on fewer products but in more geographical markets (Arora and Gambardella 1998). The restructuring took a heavy toll of incumbents; and many familiar names such as Hoechst, Union Carbide, Ciba-Geigy, Sandoz, and American Cynamid have vanished.

A number of interesting themes emerge, some of which have been studied by economists. Others remain as potentially rich veins to be mined.

International competition Why did British firms fail to exploit the rich potential of organic chemistry despite a head start, access to cheap inputs (coal tar) and to the British textile industry, and a well-functioning capital market? Many explanations, none entirely persuasive, have been offered, including the alleged bias of the British financial system towards low risk-projects (Da Rin 1998), the weak links between English universities and industry (Murmann and Landau 1998), and inferior management (Chandler 2005).

Patents Overenthusiastic patent protection in the 1870s nearly killed the French dyestuff industry, while German firms strategically used patent protection (Arora 1997). The confiscation of German patents and industrial property in Britain, France and the United States after both world wars was a setback to German firms but proved insufficient for the Americans and British to catch up. Systematic analysis of this natural experiment can shed light on the role of patents in shaping oligopolistic competition.

Markets for technology Arrow (1962) observed that Du Pont appeared to have profited as much from innovations it had licensed from others as from its own products, perhaps reflecting imperfections in the market for technology. Yet technology licensing has been extensive in chemicals (Arora et al. 2001). The market for technology dramatically changed industry structure, with accumulated production experience of incumbents insufficient to deter successful entry (Lieberman 1989).

Complementarities and industrial convergence After the Second World War, oil refining and the production of synthetic fibres and plastics came to share a common technical base. The convergence led to vertical integration by oil firms into chemicals and chemical firms into petrochemicals (Lieberman 1991). Thanks to a market for petrochemical technology, the European chemical industry was able to switch to petrochemicals

very rapidly, despite very substantial investments in coal-based technologies.

Division of labour and vertical industry structure Specialized engineering firms, which arose to provide plant construction and design services to chemical firms, led the way in diffusing petrochemical technologies worldwide (Freeman 1968). This competition prodded even large chemical firms such as Union Carbide to give licences to others, further diffusing technology and promoting entry (Arora et al. 2001). The chemical industry thus provides a clear example of the benefits of vertically disintegrated industry structures in promoting entry and competition.

The enduring lesson of the history of the chemical industry for economists is the important role of firms – their history and their capabilities – which largely explains why some countries dominated the industry for such long periods. But that history is also a strong reminder to that, in the end, even the mightiest firms must eventually bow to market forces.

See Also

- ▶ Intellectual Property, History of
- ▶ Patents
- ▶ Technical Change
- ▶ Vertical Integration

Bibliography

Arora, A. 1997. Patent, licensing and market structure in the chemical industry. *Research Policy* 26: 391–403.

Arora, A., and A. Gambardella. 1998. Evolution of industry structure in the chemical industry. In *Chemicals and long-term economic growth*, ed. A. Arora, R. Landau, and N. Rosenberg. New York: John Wiley and Sons.

Arora, A., A. Fosfuri, and A. Gambardella. 2001. *Market for technology: The economics of innovation and corporate strategy*. Cambridge, MA: MIT Press.

Arrow, K. 1962. Comments on case studies. In *The rate and the direction of inventive activity: Economic and social factors*, ed. R. Nelson. Princeton: Princeton University Press.

Chandler, A. 2005. *Shaping the industrial century: The remarkable story of the evolution of the modern chemical and pharmaceutical industries*. Cambridge, MA: Harvard University Press.

Da Rin, M. 1998. Finance and the chemical industry. In *Chemicals and long-term economic growth*, ed. A. Arora, R. Landau, and N. Rosenberg. New York: John Wiley and Sons.

Freeman, C. 1968. Chemical process plant: Innovation and the world market. *National Institute Economic Review* 45(1): 29–51.

Hounshell, D., and J. Smith. 1988. *Science and Strategy: Du Pont R&D, 1902–1980*. Cambridge: Cambridge University Press.

Lieberman, M. 1989. The learning curve, technology barriers to entry, and competitive survival in the chemical processing industries. *Strategic Management Journal* 10: 431–447.

Lieberman, M. 1990. Exit from declining industries: 'Shakeout' or 'stakeout'? *RAND Journal of Economics* 21: 538–554.

Lieberman, M. 1991. Determinants of vertical integration: An empirical test. *Journal of Industrial Economics* 39: 451–466.

Murmann, P., and R. Landau. 1998. On the making of comparative advantage: The development of the chemical industries in Britain and Germany since 1850. In *Chemicals and long-term economic growth*, ed. A. Arora, R. Landau, and N. Rosenberg. New York: John Wiley and Sons.

Chenery, Hollis B. (1918–1994)

Shantayanan Devarajan

Abstract

Hollis Burnley Chenery was born in Richmond, Virginia, in 1918. He received his Ph.D. at Harvard University, worked for the Marshall Plan in Europe, taught at Stanford University, served as Assistant Administrator of the US Agency for International Development before joining the World Bank in 1970 for a distinguished, 13-year career there. He returned to Harvard as a professor in 1983. He died in 1994.

Keywords

Chenery, H. B.; Comparative advantage; Computable general-equilibrium models; Development economics; Development strategy; Economic development; Foreign aid; Poverty

alleviation; Total factor productivity; World Bank

JEL Classifications
B31

Hollis Burnley Chenery was the consummate development economist. He defined the contours of the field with his ground-breaking research on patterns of development and development strategy. He developed tools that helped translate research into policy, and, as Vice-President for Development Policy at the World Bank, he helped shift the focus of development economics from a narrow one of economic growth to the alleviation of poverty.

Patterns of Development

In the tradition of Kuznets and Denison, Chenery was interested in how economies grow, whether there were systematic patterns in the process of development. His 1960 paper in the *American Economic Review,* 'Patterns of Industrial Growth', grew into a decade-long research project with Moshe Syrquin culminating in their 1975 book, *Patterns of Development, 1950–1970.* Many of the patterns that Chenery and Syrquin found are received wisdom today: as countries grow, the share of agriculture in GDP declines, and the shares of industry and services increase; and overall GDP growth is typically accompanied by an increase in total factor productivity (TFP) growth. Chenery and Syrquin were the first to document these patterns, using the statistical techniques available at the time, for a large number of countries in the modern era. Their work has led to Chenery–Syrquin 'norms' (interestingly, a word they never used) whereby countries could benchmark their progress in the development process. They were also aware of the limitations of this approach, identifying for example the differences between large countries and small ones, work that has been extended by Perkins and Syrquin (1989). The observed pattern of TFP growth has been questioned by, among others, Young (1995) and is still a topic of vigorous debate.

Development Strategy

In contrast with the recent work on cross-country growth (see Barro 1991), the *Patterns* work was silent on what countries could do to grow faster. Chenery answered this question in a series of major pieces on development strategy. He entered the debate between outward- and inward-looking development strategies in his 1961 *American Economic Review* paper, 'Comparative Advantage and Development Policy'. While countries should only produce those goods in which they have a comparative advantage, Chenery conjectured that comparative advantage in certain goods could be developed through careful investment policies. Chenery's notions saw a resurgence in the 1980s in the Brander and Spencer (1985) and other models of policy-induced comparative advantage. Of course, policies to create comparative advantage have to be carefully designed, especially because public investment has economy-wide impacts, as Chenery showed in his 1959 book with Peter Clark, *Interindustry Economics.*

Chenery's thinking on development strategy evolved over time. He became convinced that a country's underlying economic structure – the functioning of its labour and capital markets, its resource endowments – influenced the choices it could make in trying to create 'dynamic comparative advantage'. Using case studies, cross-country analysis and model-based analysis, he distilled this work in his 1984 book with Sherman Robinson and Syrquin, *Industrialization and Growth: A Comparative Study.*

Structure also determines how foreign aid affects the economy, as Chenery showed in his 'two-gap' model (see Chenery and Strout 1966; Chenery and Bruno 1962). *Ex ante,* an economy may be foreign-exchange-constrained or fiscally constrained. Since foreign aid is both foreign exchange and resources to the government, its impact depends on which constraint is binding. An extended version of this simple model became the workhorse model of aid agencies such as the World Bank. It saw a resurgence during the debt crises of the 1980s. It has also been criticized for neglecting the role of prices and incentives (see Easterly 1999), although it can be shown that, as long as domestic and foreign capital are imperfect

substitutes, most of the results of the two-gap model survive in a fully specified, intertemporal, general-equilibrium model.

Tools

Building on his work on the interdependence of investment decisions, Chenery and his collaborators pioneered the development of multisectoral models for investment planning, collected in his co-authored book, *Studies in Development Planning*. This work saw applications in various planning agencies, notably in India. Recognizing the limitations of linear programming approaches, Chenery encouraged the development of computable general-equilibrium (CGE) models at the World Bank and in universities. Today, CGE models are commonly used to inform policy in developing and developed countries, although they too have their limits (see Devarajan and Robinson 2005).

Redistribution with Growth

Arriving at the World Bank in 1970, Chenery proceeded to establish the first, and eventually one of the most influential, research programmes in economic development. In addition to producing academic-quality research, Chenery's group helped shape Bank policies. In 1974, Chenery and his associates published *Redistribution with Growth*, a seminal book that, while recognizing the need for direct action to alleviate poverty (especially since the high growth of the 1960s had not significantly reduced poverty), showed that wealth redistribution can and should be consistent with the promotion of economic growth. Chenery's approach has been the leitmotif of the World Bank's (and indeed most development agencies') strategy since then.

See Also

- ▶ Development Economics
- ▶ Economic Growth, Empirical Regularities in
- ▶ Foreign Aid
- ▶ Redistribution of Income and Wealth
- ▶ Structural Change
- ▶ World Bank

Selected Works

1959. (With P. G. Clark.) *Interindustry economics.* New York: Wiley.
1960. Patterns of industrial growth. *American Economic Review* 50: 624–654.
1961. Comparative advantage and development policy. *American Economic Review* 51: 18–51.
1962. (With M. Bruno.) Development alternatives in an open economy: The case of Israel. *Economic Journal* 72: 79–103.
1966. (With A. Strout.) Foreign assistance and economic development. *American Economic Review* 56: 679–733.
1971. *Studies in development planning.* Cambridge, MA: Harvard University Press.
1974. (With M. Syrquin.) *Patterns of development, 1950–1970.* Oxford: Oxford University Press.
1974. *Redistribution with growth.* London/New York: Oxford University Press.
1984. (With S. Robinson and M. Syrquin.) *Industrialization and growth: A comparative study.* New York: Oxford University Press.

Bibliography

Barro, R. 1991. Economic growth in a cross-section of countries. *Quarterly Journal of Economics* 106: 407–443.
Brander, J.A., and B.J. Spencer. 1985. Export subsidies and international market share rivalry. *Journal of International Economics* 18: 83–100.
Devarajan, S., and S. Robinson. 2005. The influence of computable general equilibrium models on policy. In *Frontiers in applied general equilibrium modeling*, ed. T. Kehoe, T.N. Srinivasan, and J. Whalley. Cambridge: Cambridge University Press.
Easterly, W.R. 1999. The ghost of financing gap: Testing the growth model of international financial institutions. *Journal of Development Economics* 60: 423–438.
Perkins, D.H., and M. Syrquin. 1989. Large countries: The influence of size. In *Handbook of development economics*, ed. H.B. Chenery and T.N. Srinivasan. Amsterdam: North-Holland.
Young, A. 1995. The tyranny of numbers: Confronting the statistical realities of the East Asian growth experience. *Quarterly Journal of Economics* 110: 641–680.

Cherbuliez, Antoine Elisée (1797–1869)

R. F. Hébert

Swiss lawyer and economist, Cherbuliez was born in Geneva into a family of French Protestants who were uprooted by the Edict of Nantes. Trained in law, Cherbuliez held a judgeship until 1835, at which time he succeeded Pellegrino Rossi at the University of Geneva as professor of public law and political economy. He also served in the Swiss Constituent Assembly and the Grand Council, but after the fall of the Conservative Republican Party in 1848 he moved to Paris and became a naturalized French citizen. A short time later, however, he returned to his homeland as professor of political economy at the University of Lausanne, preceding Léon Walras in the position. Concurrently, he held the chair of political economy at the University of Zurich from 1855 until his death in 1869.

As an economist, Cherbuliez produced nothing original, but he excelled in exposition. His writings represent a kind of mature classicism. The diadem in a collection of sparkling gems is his *Précis de la science économique et ses practicales applications*, a masterpiece of erudition, described by Schumpeter (1954, p. 501) as 'one of the best textbooks of "classic" economics'. Luigi Cossa (1880) put it on a par with Mill's *Principles*, judging it 'possibly superior'. J.E. Cairnes, in quiet affirmation of Cossa's judgement, followed Cherbuliez in his reformulation of the classical theory of value.

Cherbuliez wrote for the *Dictionnaire d'économie politique* and for the *Journal d'économie politique*, on such topics as socialism (which he opposed), charity, transportation, money and banking, taxation (he accepted Canard's theory), entrepreneurship, economic history and the history of economic thought. If he was an apostle at all, he followed Say and Bastiat. Like the former, he partitioned economics systematically into 'theory' and 'practice'. Like the latter, he wrote pamphlets in support of liberalism and the deductive method. Despite the clarity of his style and exposition, however, Cherbuliez was not widely read. He left no distinctive imprint on French economics, nor is he remembered by most textbooks in the history of economic thought.

Selected Works

1840. *Riche ou pauvre*. Paris: A. Cherbuliez.
1848. *Le Socialisme c'est la barbarie*. Paris: Guillaumin.
1862. *Précis de la science économique et de ses principales applications*, 2 vols. Paris: Guillaumin.

References

Cossa, L. 1880. *Guide to the study of political economy*. Trans. from the 2nd Italian edn. London: Macmillan.
Gide, C., and C. Rist. 1949. *A history of economic doctrines*. Trans. R. Richards. Boston: D.C. Heath.
Schumpeter, J.A. 1954. In *History of economic analysis*, ed. E.B. Schumpeter. New York: Oxford University Press.

Chernyshevskii, Nikolai Garilovich (1828–1889)

M. Falkus

Nikolai Chernyshevskii was born in Saratov in 1828 and died there in 1889. He was one of a group of 'revolutionary democrats' which included Herzen and Belinskii among its number, and Chernyshevskii became the group's outstanding intellectual leader during the critical decade following the accession of Tsar Alexander II in 1855. His greatest period of activity thus came in the aftermath of the European revolutionary upheavals in 1848–9, and coincided with Russia's defeat in the Crimean War (1853–6), the debate leading up the Emancipation of the Serfs (1861), and the subsequent post-emancipation reaction and gathering of revolutionary sentiment.

Chernyshevskii's influence both on his contemporaries and on later generations of Russian revolutionaries was profound. He undoubtedly

influenced both Marx and Lenin, and Lenin admired Chernyshevskii more than any other non-Marxist revolutionary writer. Chernyshevskii therefore holds an honoured place in Soviet literature on the development of socialist thought, and he is viewed as the main precursor of Marxism in Russia. His ideas also helped prepare the way for the development of Russian Populism (the Narodnik movement) in the 1870s.

Several major threads run through Chernyshevskii's voluminous and wide-ranging writings. He was an admirer of Western achievements and believed that Russia must modernize in order to catch up with the more civilized West. Thus although he was opposed to capitalism, he was by no means opposed to industrialization and urbanization. In this he was at odds both with the Slavophils and with the later *Narodniki*. He was a strong believer in individualism and in the benefits of enlightened self-interest. He believed that societies, shorn of such oppressive institutions as autocracy and serfdom, or of exploitative capitalism, could grow towards socialism in a rational manner. His ideas were rooted also in a belief in the Laws of History, and in the 'necessary' transition from one stage of development to another. But he also believed that societies, like individuals, could progress by revolutionary means at an accelerated pace. In particular, he argued that Russia could progress to a socialist state without having to undergo a period of capitalism (a viewpoint unacceptable to Leninists, who argued that Russia already was a capitalist country). Chernyshevskii was thus the first Russian writer to put forward a theory of accelerated social change.

In so far as Chernyshevskii perceived that Russia might 'skip' a stage of historical development by virtue of her backwardness, he may be considered a precursor of the 'concept of relative backwardness' later elaborated by Alexander Gerschenkron. In particular, Russia could take advantage of her backwardness by borrowing both institutional forms and technology from the more developed West without incurring the costs of the pioneer. But Chernyshevskii failed to develop the concept, and, as Gerschenkron has pointed out, added little to the ideas already advanced by Herzen. Indeed, in Gerschenkron's opinion 'it is not clear at all that Chernyshevskii made any independent contribution to economic analysis' (Gerschenkron 1962, p. 171).

Chernyshevskii came from a humble background. He was the son of a poor village priest and at first trained for the priesthood himself, attending the theological seminary at Saratov between 1842 and 1845. His literary and linguistic skills took him to St Petersburg, where he graduated from the department of history and philology in 1850. It was during these years that he became a radical and a revolutionary, influenced profoundly by the revolutions of 1848–9 and the debate over serfdom in Russia. His ideas were influenced by Herzen and Belinskii and also by the German philosophers (especially Feuerbach and Hegel), by French utopian socialists and by English political economists (especially Ricardo and Mill).

Following two years teaching in his native Saratov, Chernyshevskii returned to St Petersburg in 1853 and in the following year joined the staff of the literary journal *Sovreminnik* (The Contemporary). Between 1854 and 1857 he wrote most of his literary criticism, using this as a vehicle to expound his social views. He rejected any concept of 'art for art's sake', arguing the essentially political nature of aesthetics. From 1857 Chernyshevskii devoted himself almost entirely to political and social issues and wrote a series of major articles. These included *Capital and Labour* (1859), *A Critique of Philosophical Prejudices Against the Communal Ownership of Land* (1858) and *The Anthropological Principle of Philosophy* (1860). He also translated John Stuart Mill's *Principles of Political Economy* in (1860) and wrote a lengthy critique of Mill's theories.

In these and other works Chernyshevskii criticized the workings of liberal capitalism, which he condemned for its exploitation of the masses and for periodic economic crises. From Mill's wage-fund theory Chernyshevskii drew a theoretical demonstration of the inevitability of mass poverty under capitalism. He thought that under capitalism the division of labour would inevitably decrease wages, as each operation would require less skill and training and therefore would be rewarded less. And he drew the conclusion that industry, while not to be avoided, must have a different form of social organization. Social evils sprang ultimately from

poverty, and 'whoever says "poverty of the people" also says "the government is bad"'.

Chernyshevskii was a strong defender of the peasant village commune (*mir*) both on social and economic grounds. He viewed the village commune as a possible bridge which would enable Russia to avoid the capitalist stage of development, and he argued that the commune should be modernized along rational lines and become similar to workers' associations in western Europe. He attacked vehemently the terms of the Emancipation, arguing that the peasants should be given their land without having the obligation to pay redemption taxes.

Long under suspicion for propagating revolutionary views (though their message was always subtly disguised in order to escape censorship), Chernyshevskii was arrested on a pretext in July 1862 and imprisoned in the Fortress of St Peter and Paul. He was charged with 'plotting the overthrow of the existing order', and after a trial lasting two years, was ultimately sentenced to seven years' hard labour and exile for life in Siberia.

During his imprisonment Chernyshevskii continued to write a number of important works, including his most famous and influential novel *Chto Delat*' (*What is to be Done*?) (1862–3). The heroes of the novel were the 'new radicals', guided by rational self-interest ('egoism') rather than irrational beliefs. The message of the book was an optimistic one – much could be achieved by individuals who were guided by sound principles even though living in a corrupt society. The novel was serialized in *The Contemporary* (due to an oversight by the censor) and it had an immediate and profound impact on the Russian intelligentsia.

After his exile in 1864 Chernyshevskii's productive life was virtually over. He remained in Siberia – despite vain attempts by radical groups to free him – until 1883, when he was allowed to live in Astrakhan under police supervision. But only in 1889, shortly before his death, was he allowed to return to his native Saratov.

There is no doubt that Chernyshevskii's honoured place in Soviet histories of revolutionary thought owes much to the approval given his writings by Marx, Engels and Lenin. This in turn may be explained in part by Chernyshevskii's championship of the masses, his emphasis on historical forces and his materialism, as well as his own humble origins and his suffering at the hands of the Tsarist authorities. If his writings seem turgid and sometimes coarse, he was nonetheless a powerful and original thinker who did much to adapt the various strands of Western European political economy and philosophy to Russian conditions. He influenced a generation of Russian revolutionaries and did much to prepare the ground in Russia for the *Narodnik* movement of the 1870s and, later, for the spread of Marxism. There is no evidence that Chernyshevskii himself was influenced by Marx, or even that he had read Marx's works. It is probable, though, that he had read the *Communist Manifesto*, and the first volume of *Das Kapital* was sent to him in Siberia in 1872.

Selected Works

1853. *Selected philosophical essays*. Moscow: Foreign Languages Publishing House.

1862–3. *A vital question: Or what is to be done?* (trans: Dale, N.H. and S.S. Skidelsky). New York: T.Y. Crowell & Co., 1886.

References

Gerschenkron, A. 1962. *Economic backwardness in historical perspective*. Cambridge, MA: Harvard University press.

Pereira, N.G. 1975. *The thought and teachings of N.G. Chernyshevskii*. The Hague: Mouton.

Randall, F.B. 1967. *N.G. Chernyshevskii*. New York: Twayne Publishers.

Chevalier, Michel (1806–1879)

P. Bridel

Keywords

Chevalier, M.; Free trade; Leroy-Beaulieu, P. P.; Liberalism; Rossi, P.L. E.; Saint- Simon, C. H.; Walras, L

JEL Classifications
B31

Born in Limoges, 13 January 1806; died in Paris, November 1879. Undoubtedly one of the most eminent 19th-century French economists, Chevalier belongs to that most typical brand of engineer-economists. First in his class (*major*) at the Ecole Polytechnique in 1830 and member of the Corps des Mines as an economist, Chevalier came very early under the spell of Saint-Simon's utopian doctrine. From his early editorship of the Saint-Simonian newspaper *Le Globe* (1830–2) and his subsequent sentence to a year in jail (for 'outrage to morals' for publishing advanced ideas on the liberation of women, sexual liberty and the need for communal life) to a made-to-measure niche as economic adviser to Napoleon III and 'éminence grise' to the Second Empire business and banking establishment, Chevalier applied his brilliant mind to various current problems and policy issues without managing, however, to escape completely from the Saint-Simonian mystique. His main claim to fame, the Anglo-French Treaty of 1860 (the Cobden–Chevalier Treaty), an important if short-lived interruption in the general protectionist policy of France, is one of the best illustrations of these twin components of Chevalier's approach to economics and economic policy: weak on the analytics and very strong on the factual analysis with a touch of Saint-Simonian idealism.

Together with public works, cheap bank credit and education, free trade is one of the articles of faith he took over from the Saint-Simonian doctrine. Chevalier returned to these issues throughout his life (notably in his penetrating analysis of the American economy and banking system in the early 1830s which earned him later the nickname of 'Economic Tocqueville'). Binding these various elements with a quasi-philosophical concept of association (as the cornerstone of social order), Chevalier suggests a broad theory of economic growth which he considered flexible enough to be applied to different times and countries.

His Saint-Simonian antecedents and his extensive travelling (to England, Egypt and foremost to the United States) rendered Chevalier suspicious of all 'absolutist' economic theory. In fact, in his most technical chapters (particularly on money) Chevalier never digs beneath the surface of things and contributes very little, if anything, to analytic economics. His only systematic work, his *Cours* (1843; 1844; 1850) delivered at the Collège de France offers little more in the field of theory than a lengthy (and flat) apology for Say's brand of 'vulgar' liberalism. With Rossi, his predecessor, and Leroy-Beaulieu, his successor at the Collège de France, Chevalier was in fact largely responsible for introducing and perpetuating in academic circles the liberal orthodoxy that was to bar Walras from getting an appointment in the 1860s and that dominated French economics for so long that as late as 1939 Keynes could still quip about its lack of 'deep roots in systematic thought' (1939, p. xxxii).

Selected Works

1843, 1844, 1850. *Cours d'économie politique*. 3 vols. Paris: Capelle.

Bibliography

Keynes, J.M. 1939. French Preface to *The general theory of employment, interest and money*. In *Collected writings*. vol. 7, 31–35. London: Macmillan, 1973.

Walch, J. 1975. *Michel Chevalier, économiste, saint-simonien*. 1806–1879. Paris: Vrin (with extensive bibliography).

Cheysson, Jean-Jacques Emile (1836–1910)

R. F. Hébert

French engineer, economist and statistician, Cheysson was born in Nîmes and died in the Swiss Alps. Schooled at the Ecole Polytechnique and the Ecole des Ponts et Chaussées, he served with distinction in the Corps of Civil Engineers,

demonstrating his ingenuity during the German siege of Paris (1870) by converting train stations to flour mills (using locomotive engines as the power source), thereby increasing bread production. Only when wheat supplies were eventually exhausted did the city finally capitulate. After the armistice, Cheysson became factory director at Creusot, the huge industrial complex that was bombed during World War II, where he immersed himself in the microeconomics of the firm and began to develop an analytical programme which anticipated the main lines of what we now call econometrics.

Calling his method 'geometric statistics', Cheysson presented its outline to the Paris Statistical Society in 1885 as a scientific approach to 'the practical solution of business problems'. His technique rested on the twin pillars of theory and observation. It combined the spirit of Cournot's economics with attention to recorded data, using geometry to display concrete facts and to interpolate gaps in available statistics. In spirit and scope, but considerably ahead of its time, it mirrored the objectives of the Econometric Society, established at Lausanne in 1931.

A potential alliance between Cheysson and Léon Walras eventually soured, thus cutting off what might have been a productive channel of communication for Cheysson's new method. Cheysson taught geometric statistics to his students at the Ecole des Mines and the Ecole des Sciences Politiques (he held the first chair of economics at each institution), but he inspired no group of followers the way Walras or Marshall did, and the powerful originality of his contribution gradually faded, only to be rediscovered in the present century by Staehle (1942, p. 322) and Schumpeter (1954, p. 842), and reconstructed piecemeal by Hébert (1972, 1973, 1974).

Within the narrow compass of 35 pages Cheysson enriched the theories of statistical demand; revenue and cost curves; profit maximization; spatial market boundaries for raw materials and finished products; wages; product and quality variation; investment; and taxation. His deft handling of these difficult subjects while the titans of economic theory were debating the psychological premises of value theory, constitutes a remarkable performance, even by modern standards.

After 1890 Cheysson turned his energies increasingly towards that branch of ideas that the French call 'social economy'. A follower of LePlay since their first meeting in 1864, Cheysson shared his colleague's interest in social and economic reform. LePlay's school emphasized moral and religious considerations in the economic order, especially the primacy of the family, the rights of workers and the duties of employers. Under Cheysson's leadership, the Société d'Economie Sociale (founded by LePlay in 1856) wedged itself between the socialists on the left and the liberals on the right. Unwilling to accept the evils of poverty and the misfortunes of the workers, yet rejecting socialist remedies, LePlay's school sought amelioration through the encouragement of private initiative. They considered social reform as much a matter of economics as morality. Cheysson was thrice president of the Société d'Economie Sociale, and was elected to the Académie des Sciences Morales et Politiques in 1901. He left behind a literary legacy that numbered over 500 publications, embracing such diverse topics as economics, statistics, geography, agriculture and social hygiene.

Selected Works

1911. *Oeuvres choisies*. 2 vols. Paris: A. Rousseau.

Bibliography

Colson, L.C. 1913. Notice sur la vie et les travaux de M. Emile Cheysson. Académie des Sciences Morales et Politiques. *Séances et travaux* 179: 153–187.

Hébert, R.F. 1972. A note on the historical development of the economic law of market areas. *Quarterly Journal of Economics* 86(November): 563–571.

Hébert, R.F. 1973. Wage cobwebs and cobweb-type phenomena: An early French formulation. *Western Economic Journal* 11(December): 394–403.

Hébert, R.F. 1974. The theory of input selection and supply areas in 1887: Emile Cheysson. *History of Political Economy* 6: 109–113.

Schumpeter, J.A. 1954. In *History of economic analysis*, ed. E.B. Schumpeter. New York: Oxford University Press.

Staehle, H. 1942. Statistical cost functions: Appraisal of recent contributions. *American Economic Review* 32(June): 321–333.

Chicago School

M. W. Reder

Abstract

This article deals with the history and main protagonists of the Chicago School from c. 1930 to 1985. The two main beliefs of members of the School are (*a*) that neoclassical price theory can explain observed economic behaviour, and (*b*) that free markets efficiently allocate resources and distribute income, implying a minimal role for the state in economic activity. Chicagoans maintain that no opportunity for arbitrage gains goes unexploited, and subscribe to the efficient markets hypothesis. Their 'disciplinary imperialism' leads them frequently to challenge conventional wisdom by applying price theory to seemingly non-economic topics.

Keywords

Aggregate demand; Arbitrage; Axiomatic theories; Becker, G.; Buchanan, J.; Capital asset pricing model; Capital theory; Chicago school; Clark, J.; Coase theorem; Coase, R.; Competition; Cost of time; Cowles Commission; Deregulation; Director, A.; Dividends; Douglas, P.; Education; Efficient markets hypothesis; Egalitarianism; Family, economic analysis of; Finance; Free markets; Friedman, M.; Friedman, R.; General equilibrium theory; Hardy, C.; History of economic thought; Human capital; Imperfect competition; Income distribution; Industrial organization; Institutionalism; International trade; Keynesianism; Knight, F.; Laissez faire; Lange, O.; Laughlin, J.; Law, economic analysis of; Leland, S.; Lerner, A.; Lewis, H.; Lucas, R.; Market prices; Market socialism; Marschak, J.; Mathematical economics; Metzler, L.; Miller, M.; Millis, H.; Mitchell, W.; Modigliani, F.; Monetarism; Monetary theory; Money supply; Monopolistic competition; Muth, J.; Natural monopolies; Nef, J.; Neoclassical monetary theory; Neoclassical price theory; New Deal; Pareto optimality; Political economy; Political economy of policy reform; Political markets; Positive economics; Posner, R.; Price theory; Progressive and regressive taxation; Public ownership; Quantitative techniques; Racial discrimination; Rational expectations; Reder, M. W.; Regulation; Reserve/deposit ratio; Resource allocation; Risk; Schultz, H.; Schultz, T.; Simons, H.; Social costs; Statistical demand curves; Statistics and economics; Stigler, G.; Stock prices; Testing; Trade unions; Transaction costs; Uncertainty; Veblen, T.; Viner, J.; Wallis, W.; Wright, C.

JEL Classifications

B5

To identify a Chicago School of economics requires some demarcations, both of ideas and persons, that may not be universally accepted. Justification for these decisions must be heuristic; that is, they facilitate the story to be told. But it is not denied that there may be alternative accounts that would entail different demarcations. In this account, the 'Chicago School' is and has been centred in the University of Chicago's Economics Department from about 1930 to the present (1985). However, it is convenient to define the School so as to include many members of the large contingent of economists in the Graduate School of Business and the group of economists and lawyer-economists in the Law School. Largely because of the intellectual loyalty of former students, the influence of the Chicago School extends far beyond the University of Chicago to the faculties of other universities, the civil service, the judiciary and private business. Moreover, this influence is not confined to the United States.

To restrict the retrospective horizon of the School to 1930 implies exclusion of a number of famous economists who had been on the University of Chicago faculty before that time; for example, Thorstein Veblen, Wesley C. Mitchell, J.M. Clark, J. Laurence Laughlin, C.O. Hardy. However, none of these shared the intellectual

characteristics that have typified members of the Chicago School as defined here.

In a nutshell, the two main characteristics of Chicago School adherents are: (1) belief in the power of neoclassical price theory to explain observed economic behaviour; and (2) belief in the efficacy of free markets to allocate resources and distribute income. Correlative with (2) is a tropism for minimizing the role of the state in economic activity.

Before discussing these characteristics in detail, let me give a brief historical account in which it is convenient to divide the history of the School into three periods: (1) a founding period, in the 1930s; (2) an interregnum, from the early 1940s to the early 1950s; and (3) a modern period, from the 1950s to the present.

During the founding period, the Chicago Economics Department contained a wide diversity of views both on methodology and public policy. Institutionalist views were well represented among the senior faculty, and institutionally oriented students constituted a large part of the graduate student population. Among the prominent Institutionalists were the labour economists H.A. Millis and (one side of) Paul H. Douglas; the economic historians John U. Nef and C.W. Wright, and Simeon E. Leland, a Public Finance specialist and long-time department chairman.

Like other social science departments at Chicago, economics was actively engaged in developing the (then) embryonic 'quantitative techniques'. The leading figures in quantitative methods were Henry Schultz, a pioneer student of statistical demand curves, who taught the graduate courses in mathematical economics and mathematical statistics, and Paul Douglas who was (during the 1920s and 1930s) a leader in the estimation of and the measurement of real wages and living costs.

However, it is generally agreed that the progenitors of the Chicago School were Frank H. Knight and Jacob Viner. These two scholars shared an intense interest in the history of economic thought and both were, broadly speaking, devotees of neoclassical price theory. However, their intellectual styles and temperaments were quite different, and their personal relations were not close. Apart from his interest in the history of thought, Viner was primarily an applied theorist working on problems in international trade and related issues in monetary theory. Knight's work was focused on the conceptual underpinnings of neoclassical price theory, and his main concerns were to clarify and improve its logical structure.

Temperament and intellectual focus combined to make Knight a formidable critic, both of ideas and their protagonists. This led to a good deal of friction between him and both Douglas and Schultz. Personalities aside, Knight was strongly averse to the quantification of economics and was very outspoken on this, as on most other matters (For further details, see Reder 1982, pp. 362–5).

By contrast, Viner was rather sympathetic to the aspirations of 'quantifiers', though sceptical of their prospects for success, at least in the near future. Viner's sympathy for quantitative work was prompted by the strong empirical bent of his own research, although friendship for Douglas and Schultz may also have been involved. On the other hand, Knight's purely theoretical studies of capital theory, risk, uncertainty, social costs, and so on, generated neither need for empirical verification nor exposure to research that might have offered it. As a result, Knight's relations with Douglas and Schultz were ridden with conflict, and theoretical disagreements with Viner spilled over into barbed comments to graduate students and kept personal relations (between Knight and Viner) from becoming more than merely correct (Reder 1982, p. 365).

What Knight and Viner had in common was a continuing adherence to the main tenets of neoclassical price theory and resistance to the theoretical innovations of the 1930s, Monopolistic Competition and Keynes's *General Theory.* This theoretical posture paralleled an antipathy to the interventionist aspects of the New Deal and the full employment Keynesianism of its later years. Viner, who was actively consulting the government throughout the period, was much less averse to New Deal reforms than Knight and his protégés. However, there was a sharp contrast between the views of Knight and Viner, on the one hand, and those of avowed New Deal

supporters such as Douglas, Schultz and some of the Institutionalists.

As a result of the division of faculty views, on both economic methodology and public policy, the graduate student body was exposed to a diversity of thought patterns and did not exhibit a great degree of conformity to any particular one. But despite their many disagreements, an effective majority of the Chicago faculty concurred in a set of degree requirements (for the PhD) that stressed competence in the application of price theory. These requirements were quite unusual in the 1930s and the process of satisfying them exercised a great influence in forming a (common) view of the subject among the students, in which price theory was of major importance.

The most important of the requirements was that all PhD candidates, without exception, pass preliminary examinations in both price theory and monetary theory. These examinations were difficult and attended with an appreciable failure rate. Even on second and third trials, there was a non-negligible probability of failure, with the result that some students were (and are) unable to qualify for the doctorate. For most students, the key to successful performance on the examinations was mastery of the material presented in relevant courses, especially the basic price theory course (301) and study of previous examinations.

For over half a century, the need to prepare for course and preliminary examinations, especially in price theory, has provided a disciplinary–cultural matrix for Chicago students. Examination questions serve as paradigmatic examples of research problems and 'A' answers exemplify successful scientific performance. The message implicit in the process is that successful research involves identifying elements of a problem with prices, quantities, and functional relations among them as these occur in price theory, and obtaining a solution as an application of the theory.

Although the specific content of examination questions has evolved with the development of the science, the basic paradigm remains substantially unchanged: economic phenomena are to be explained primarily as the outcome of decisions about quantities made by optimizing individuals who take market prices as data with the (quantity) decisions being coordinated through markets in which prices are determined so as to make aggregate quantities demanded equal to aggregate quantities supplied.

Of course, students vary in the degree to which they assimilate price theoretic ideas to their thought processes, and resistance to these ideas was probably greater in the 1930s than later. Nevertheless, regardless of their special field of interest, all students were compelled to absorb and learn to use a considerable body of economic theory. In the 1980s these skills are very widespread, but in the 1930s they were rarely found and served to distinguish Chicago-trained PhD's – especially in applied fields – from other economists.

Despite the common elements of their training, as in other institutions, doctoral students tended to identify themselves with one or another particular faculty member, usually their dissertation supervisor. Thus each of the major figures in the department was associated with a cluster of advanced students. One such cluster, associated with Knight in the mid-1930s, became of very great importance in the history of the Chicago School. Key members of this cluster were Milton Friedman, George Stigler and W. Allen Wallis. The group established close personal relations with two junior faculty members, Henry Simons and Aaron Director, who were also protégés of Knight. Another member of the group was Director's sister, Rose, who later married Milton Friedman.

It was this group that provided the multigenerational linkage in intellectual tradition that is suggested by the term 'Chicago School'. Although they admired Knight, and were devoted to him, the intellectual style of Friedman, Stigler, et al. was very different from Knight's. They were thoroughgoing empiricists with a distinct bias toward application of quantitative techniques to the testing of theoretical propositions. In their empirical bent and concern with 'real world' problems, they were much closer to Viner than to Knight, but, whatever the reason, they identified with the latter.

Partly because of his important role in the teaching of theory to undergraduates and (less well-prepared) beginning graduates, in the 1930s and until his untimely death in 1946, Henry

Simons exercised an important influence on Chicago students. But he is remembered mainly for his essays on economic policy (collected in Simons 1948) which constituted the principal statement of Chicago laissez-faire views during this period.

Simons's view had a distinctly populist flavour that is absent from those more recently associated with Chicago economics. For example, he favoured use of government power to reduce the size of large firms and labour unions. Where such policies would lead to unacceptable losses of efficiency (e.g. 'natural monopolies'), Simons favoured outright public ownership. In sharp contrast to more recent Chicago statements on the matter, Simons emphatically supported progressive income taxation to promote a more egalitarian distribution of income (Simons 1938).

Finally, Simons proposed a requirement of 100 per cent reserves against demand deposits and restriction of Federal Reserve discretion in monetary policy in favour of fixed rules designed to stabilize the price level (Simons 1948). In this he was the direct forbear of Chicago monetarism, as later developed by Friedman and Friedman's students.

Historically, Friedman, Stigler and Wallis were both the intellectual and the institutional heirs of Knight and Viner. The story of Chicago economics would be less convoluted if the succession had been a matter of the older generation appointing their best students to succeed them. But it was not that simple. On the eve of World War II there was great concern, within the Economics Department and (probably) in the central administration as well, that Chicago had none of the leading figures in the new theoretical developments of the period; that is, in nonperfect competition and Keynesian macroeconomics.

To rectify this, in 1938, they appointed Oscar Lange as assistant professor. In addition to his credentials as a contributor to the literature of Keynes's *General Theory,* especially its relation to general equilibrium theory, Lange was a leading participant in the current debate on the possibility of market socialism and its (alleged) advantages relative to laissez-faire capitalism in terms of efficiency. Further, he had made a number of contributions to mathematical economics and was able to provide backup support for Henry Schultz in that subject area, and in mathematical statistics as well.

As an outspoken and politically active socialist, Lange's views were diametrically opposed to laissez faire. That he managed to stay on friendly terms with virtually all of his colleagues was a testimonial both to his own tact and to their tolerance of dissent. Of course, it was no accident that the principal socialist in the Chicago tradition should have been a *market* socialist.

Within a few months of Lange's appointment, Henry Schultz was killed in an automobile accident and Lange became the sole mathematical economist in the Chicago department. Within a year the loss of Schultz was compounded by the partial withdrawal of Douglas from academic life to pursue a political career. Still further, with the outbreak of World War II, Viner became increasingly involved in Washington and, ultimately, in 1945, he resigned to accept an appointment at Princeton.

As a result of these losses, the Department had to be rebuilt. The process of reconstruction began during the war years, with Lange taking a leading role. He was very anxious to recruit colleagues who were leaders in current theoretical developments, especially in mathematical economics. Failing to obtain his first choice, Abba Lerner, he readily accepted Jacob Marschak and, for a short period, collaborated with the latter in making further appointments both to the Department and to the Cowles Commission, which had located at the University of Chicago in 1938. The collaboration ended abruptly in 1945 when Lange resumed Polish citizenship to become ambassador to the United States and, subsequently, to fill many other high positions in the socialist government of Poland.

During the war years, T.W. Schultz was attracted from Iowa State. A leading figure in agricultural economics, Schultz soon became chairman, a position from which he exercised much influence for over two decades. In addition to Schultz, in 1946 the Department acquired Lloyd Metzler to teach international trade and a number of younger theorists and econometricians

associated mainly with the Cowles Commission. Whatever was the intention, these appointments served as a counterweight to the more or less contemporaneous appointments of Friedman (to the Economics Department) and Wallis (to the Business School).

There then ensued a struggle for intellectual pre-eminence and institutional control between Friedman, Wallis and their adherents on one side, and the Cowles Commission and its supporters on the other. The struggle persisted into the early 1950s, ending only with the partial retirement of Lloyd Metzler (due to ill health) and the departure of the Cowles Commission (for Yale) in 1953. While not monolithic, the Chicago economics department that emerged from this conflict had a distinctive intellectual style that set it apart from most others.

In positive economics, this style involves de-emphasizing the role of aggregate effective demand as an explanatory variable and stressing the importance of relative prices and 'distortions' thereof. In economic policy, it involves stressing the beneficial effects of allowing prices to be set by market forces rather than by government regulation. In an important sense, 'Chicago economics' in the 1950s and 1960s was simply an extension of the ideas of the Knight coterie of the 1930s. Indeed, some of the key figures – notably Friedman, Stigler and Wallis – of that group were leading Chicago economists in the later period as well. Moreover, they were consciously concerned with explicating the continuity of the tradition and preserving it (see below).

The close personal relations of the members of the Knight coterie, maintained for over a half century, has reinforced the strong common elements in their idea-systems and made it easy to ignore the (important) points of disagreement, both among themselves and with others. As already mentioned, Friedman, Stigler and Wallis, like most Chicago economists of their own and subsequent cohorts, believe strongly in use of statistical data and techniques for testing economic theories. In this they differ from Knight, Simons, James Buchanan, Ronald Coase (1981) and a significant minority of other economists associated with Chicago, either as graduate students or faculty, who believe (on various grounds) that the validity of an economic theory lies in its intuitive appeal and/or its compatibility with a set of axioms, rather than in the conformity of its implications with empirical observation.

A second disagreement concerns the consistency of policy advocacy in any form, with the methodology applied in positive economics (The most influential general description of this methodology is chapter 1 of Friedman 1953). This methodology recommends that explanations of economic behaviour be based on a model of (individual) decisions of resource allocation (among alternative uses) designed to maximize utility subject to the constraints of market prices and endowments of wealth. Market prices are presumed to be set so as to equate quantities supplied with those demanded, for all entities traded.

As traditionally applied by neoclassical economists with a predilection for laissez faire, this methodology coexists with advocacy of government policies designed to promote that objective. But in the late 1960s one group of Chicago economists led by Stigler (who had returned to Chicago in 1958 as Walgreen Professor in both the Economics Department and the Business School) began to apply the tools of economic analysis to the investigation of the determinants of political activity, especially government intervention in resource allocation. Thus study of the regulatory and taxing activities of the state became directed not simply at demonstrating their adverse effects upon economic efficiency, but primarily to explaining their occurrence as an outcome of the operation of 'political markets' for such activities.

So analysed, interventions traditionally viewed as efficiency impairing, such as tariffs, require reinterpretation. An individual's resources include not only his command over goods and services acquired through conventional markets, but also his political influence (however measured). Government interventions are considered to be endogenous outcomes of a political-economic process, reflecting the political as well as the economic wealth of decision making units, and not as aberrations of an exogenous state (e.g. see Stigler 1982). So viewed, criticism of political outcomes

is no more warranted than criticism of the expenditure behaviour of sovereign consumers; both are outcomes of the free choice of resource owners.

This is not to suggest that the 'political economy' wing among Chicago economists has become indifferent to laissez faire. On the contrary, opposition to government intervention (e.g. regulation) among Stigler and his allies is quite as strong as it ever has been. During the past decade many economists and lawyers at some time affiliated with the Law and Economics group at Chicago have been prominent advocates of deregulation. However, tension between advocacy of reform, and positive analysis of the political process through which reform must be achieved, presents a continuing existential problem to the heirs of the Chicago tradition. Although they are well aware of the problem, thus far they have refrained from divisive dispute and treat exercises in political advocacy as a consumption activity by those engaged.

Political science is only one of the fields into which Chicago economics has expanded during the past quarter century. Beginning in the early 1940s and accelerating in the last two decades under Richard Posner's leadership, the economic analysis of legal institutions has become an important area of research both for economists and for legal scholars. Further, using the theory of labour supply as a point of departure, the economic analysis of the family has become an important part of the study of population, marriage, divorce and family structure. This development has challenged sociological and psychological modes of explanation in fields that had long been considered provinces of these other disciplines. Still further, the theory of human capital has had a major impact on the study of education.

It is convenient to date the 'disciplinary imperialist' phase of the Chicago School as beginning in the early 1960s and continuing to the present. However, its roots go back into the 1930s; since that time there has been, at least in the oral tradition, a tropism for application of the tools and concepts of price theory to (seemingly) alien situations, and for taking delight in confronting conventional wisdom with the results. Correlatively, there has been a strong tendency to resist explanations of behaviour that do not run in terms of utility maximization by individual decision-makers coordinated by market clearing prices.

However, until well into the 1950s, the disciplinary imperialist aspect of the Chicago paradigm was overshadowed by the struggle to defend the integrity of neoclassical price theory from the attacks of Keynesians at the macro level and the attempts of various theorists of nonperfect competition to provide alternatives at the micro level. The counterattack on the *General Theory* produced a revival of neoclassical monetary theory in a refined and empirically implemented form; this revival is associated with the work of Milton Friedman (1956).

The struggle to re-establish the competitive industry as the dominant model for explaining relative prices was led by Stigler (1968, 1970), and generated much of the theoretical and empirical literature of the field of Industrial Organization. Both in Industrial Organization and Money-Macro, the earlier debates continue, with Chicago-based participants being identifiable as partisans of the standpoints of Friedman and Stigler a quarter of a century ago. However, in the 1970s and 1980s the topics related to these debates have been forced to share centre stage with newer subjects.

The expansion of Chicago economics beyond the traditional boundaries of the discipline began in the middle and late 1950s; two early examples were H.G. Lewis's application of price theory to the 'demand and supply of unionism' (Lewis 1959) and Gary Becker's dissertation on racial discrimination (Becker 1957). These were followed in the 1960s and 1970s by a number of others, as already mentioned. Many of these are more or less straightforward applications of conventional price theory to new problems. However, the analysis of time as an economic resource (Becker 1965) has led to important improvements in the theory of household behaviour.

The analysis of time is also related to a methodological tendency to reject differences in tastes (including attitudes, opinions and beliefs in 'tastes') as a source for explanations of cross-individual differences in behaviour (Stigler and

Becker 1977; Becker 1976). The rejection is based on the contention that (1) seeming differences of taste are usually reducible to differences of cost and (2) statements about cost differences are much more amenable to empirical test. While this methodological principle has met with resistance, at Chicago as elsewhere, it is reflected in a great deal of ongoing research, especially where cost of time is an important variable.

A separate path of disciplinary expansion has arisen in the field of Finance. Whether, prior to the 1960s, this field was a province of Economics, is a point that it is convenient to bypass. But unquestionably, prior to the theoretical developments initiated by Modigliani and Miller's famous paper (1958) on the (non) relation of stock prices and dividends, the theory of price. Subsequent developments have completely reversed that situation, so that in the mid-1980s, the 'capital asset pricing model' has become an integrating matrix for the theories of security prices, asset structure of the firm, and, via the study of executive compensation, wages.

The dominant idea underlying these developments is that, save for transaction costs, *on average* no opportunity for arbitrage gains goes unexploited. One implication of this is the proposition that there is 'no free lunch'; another implication is that no specifiable algorithm can be found that will enable a resource owner to utilize publicly available information to predict movements of asset prices well enough to gain by trading. The latter implication is tantamount to the 'hypothesis of efficient markets'.

While not formally identical with rational expectations, efficient markets will support any behaviour conforming to rational expectations, but will be compatible with other models of expectations only where one or another set of correlated forecast errors (across individuals) is assumed. Moreover, so long as expectations are rational, and regardless of how they are generated, there is no way in which variables operating through expectations can improve upon the neoclassical explanation of relative prices and quantities. This obviates any need for augmenting economic theory by variables reflecting psychological or sociological factors that operate upon individual decision-making via expectations. Obviously, such a theory of expectations is strongly supportive of the claims of economic theory in interdisciplinary competition.

The interrelated ideas of rational expectations and efficient markets originated at Carnegie-Mellon in the work of Muth (1961) and Modigliani and Miller (1958) rather than at Chicago. However, their consonance with the Chicago paradigm is such that they have found a home in the Chicago Business School under the leadership of Miller and his students, and (since the mid-1970s) in the Economics Department under Robert Lucas, rather than in their place of origin. While the claim of Chicago to be the primary locus for research in these fields is a strong one, it is a claim more subject to challenge than analogous claims in some other fields.

Yet a third Chicago innovation of the late 1950s is the 'Coase Theorem' (Coase 1960). In essence this theorem states that, ignoring transaction costs, if there is any reallocation of goods, claims, rights (especially property) or alteration of institutions that – after making compensating side payments to losers – increases the utility of everyone, said reallocation will occur. If rationality is a maintained hypothesis and transaction costs are negligible, the theorem becomes a tautology. Thus the empirical content of the theorem will vary inversely with the importance attributed to transaction costs, which serve as a conceptual receptacle for all forces bearing upon decision-making other than those explicitly incorporated in the theory of price. To consider the Coase Theorem empirically important is to believe that transaction costs and departures from rationality are unimportant.

Put differently, the Coase Theorem suggests that the real world tends towards a position of Pareto optimality. Of course, for given tastes and technology, there may be a different Pareto optimum for each distribution of wealth. Therefore, to the extent that the distribution of wealth is exogenous and has important behavioural consequences, the predictive implications of both Pareto optimality and the Coase Theorem are less salient. Thus the rise in influence of the Coase Theorem at Chicago has more or less

paralleled a decline in the marked concern with income distribution that existed in the 1930s and 1940s, especially in the work of Henry Simons (Reder 1982, p. 389).

When objects of exchange are taken to include legislation and other political variables, the Coase Theorem strongly suggests that the forces of decentralized decision-making that govern production and exchange also control changes in laws and institutions. Thus belief in the Coase Theorem is – or should be – conducive to political passivity. Nevertheless, not all Chicago economists are politically quiescent. But with few exceptions, they are generally conservative, though with considerable differences of shading and intensity of belief, and in taste for political controversy. Probably these differences parallel differences in the degree to which they accept economic explanations of political behaviour. Perhaps the most common characteristic of Chicago economists is distrust of the state. This distrust, together with the belief that, given time, voluntary exchange will usually generate truly desirable reforms, acts as a powerful brake on wayward impulses to improve society through political action.

The saga of the Chicago School is at once the story of the evolution of a set of ideas – a paradigm – and of a particular institution with which its leading protagonists have been associated. In this essay I have emphasized certain central theoretical ideas and historical events to the exclusion of detailed coverage of applied work and mention of the individuals responsible for it. However, it is the association of these central ideas with an identifiable, multigenerational group of individuals located at a particular institution that justifies the title of this article. Many of the key individuals in this history – Director, Friedman, Stigler, Wallis – are still alive, intellectually active and in close touch with their successors on the Chicago faculty. This continuity, both of personalities and ideas, is a distinctive feature of the intellectual tradition called the Chicago School.

In the mid-1980s the vitality of this tradition is threatened more by the growing acceptance of many of its key ideas than by resistance to them. A quarter century ago, Chicago economics was distinguished by its emphasis on the importance of competition and money supply. Arguably, in 1985, these views and their extensions have become mainstream economics, leaving the story of the Chicago School as a nearly closed episode in the history of economic thought. While such an argument may prove valid, it is too soon to tell.

See Also

- ▶ Chicago School (New Perspectives)
- ▶ Coase, Ronald Harry (Born 1910)
- ▶ Douglas, Paul Howard (1892–1976)
- ▶ Friedman, Milton (1912–2006)
- ▶ Knight, Frank Hyneman (1885–1962)
- ▶ Lange, Oskar Ryszard (1904–1965)
- ▶ Laughlin, James Laurence (1850–1933)
- ▶ Metzler, Lloyd Appleton (1913–1980)
- ▶ Schultz, Henry (1893–1938)
- ▶ Schultz, T. W. (1902–1998)
- ▶ Simons, Henry Calvert (1899–1946)
- ▶ Stigler, George Joseph (1911–1991)
- ▶ Viner, Jacob (1892–1970)

Bibliography

Becker, G.S. 1957. *The economics of discrimination*. Chicago: University of Chicago Press.

Becker, G.S. 1965. A theory of the allocation of time. *Economic Journal* 75: 493–517.

Becker, G.S. 1976. *The economic approach to human behavior*. Chicago: University of Chicago Press.

Coase, R.H. 1960. The problem of social cost. *Journal of Law and Economics* 3: 1–44.

Coase, R.H. 1981. *How should economists choose?* Washington, DC: American Enterprise Institute for Public Policy Research.

Friedman, M. 1953. *Essays in positive economics*. Chicago: University of Chicago Press.

Friedman, M. 1956. *Studies in the quantity theory of money*. Chicago: University of Chicago Press.

Lewis, H.Gregg. 1959. Competitive and monopoly unionism. In *The public stake in union power*, ed. P.D. Bradely. Charlottesville: University of Virginia Press.

Modigliani, F., and M.H. Miller. 1958. The cost of capital, corporation finance and the theory of investment. *American Economic Review* 48: 261–297.

Muth, J.F. 1961. Rational expectations and the theory of price movements. *Econometrica* 29: 315–335.

Reder, M.W. 1982. Chicago economics: Permanence and change. *Journal of Economic Literature* 20 (1): 1–38.
Simons, H.C. 1938. *Personal income taxation*. Chicago: University of Chicago Press.
Simons, H.C. 1948. *Economic policy for a free society.* Chicago: University of Chicago Press.
Stigler, G.J. 1968. *The organization of industry.* Homewood, III: Richard D. Irwin.
Stigler, G.J. 1982. *Economists and public policy.* Washington, DC: American Enterprise Institute for Public Policy Research.
Stigler, G.J., and G.S. Becker. 1977. De gustibus non est disputandum. *American Economic Review* 67 (2): 76–90.
Stigler, G.J., and J.K. Kindahl. 1970. *The behavior of industrial prices*. New York: Columbia University Press for the National Bureau of Economic Research.

Chicago School (New Perspectives)

Ross B. Emmett

Abstract

M. W. Reder's entry on the Chicago School closed with the claim that the final chapter of the School's history was about to end. Chicago economics has changed, but it has also stayed the same. Each of the four movements of recent Chicago economics are rooted in common themes of the tradition. As well, our interpretation of economics at Chicago has evidenced both continuity and change. Historians are examining the history of the institutional structure of Chicago economics, as well as the histories of specific fields at Chicago (labour, economic history, quantitative analysis) and finding both change and continuity in the tradition.

Keywords

American Economic Association; Becker, G; Chicago model; Chicago School; Coase, R. H; Cowles Commission; Director, A; Friedman, F; Griliches, Z; Hansen, L; Harberger, A; Heckman, J; Johnson, D. G; Johnson, H. G; Knight, F; Lange, O; Levitt, S; Lewis, H. G; Lucas, R; Markowitz, H; Miller, M; Mundell, R; Murphy, K; National Bureau of Economic research; Positive economics; Rees, A; Sargent, T; Schultz, G; Schultz, T. W; Stigler, G; Viner, J

JEL Classifications

B5

The history of Chicago economics remains a story of continuity and change.

M.W. Reder closed the entry on the Chicago School in the first edition of *The New Palgrave* (and reproduced in this edition) with the claim that the final chapter of the School's history was about to end. Perhaps he was right: the apex of the School's influence on public policy – the presidency of Ronald Reagan – ended in 1988. By that time key figures in the School's history had retired, become inactive, left the University of Chicago, or died. Milton Friedman retired in 1977 and moved to the Hoover Institution at Stanford University, where he was eventually joined by Aaron Director (linchpin of the early Chicago law and economics movement) and George Schultz (former dean of the University of Chicago's Graduate School of Business and Secretary of State under President Reagan); he died in late 2006. Arnold Harberger stepped down as chair of the Economics Department in the early 1980s and moved to UCLA shortly thereafter, following the previous departure of long-time graduate advisor H. Gregg Lewis to Duke in 1977. T.W. Schultz, former department chair, was largely inactive as a scholar by the late 1970s; his student and collaborator for many years, D. Gale Johnson, retired in the early 1980s. In international economics, Robert Mundell left the university in the early 1970s and Harry Johnson died in 1979. Of the early leaders, only George Stigler (industrial organization) and Ronald Coase (law and economics) remained active at Chicago, although both were retired.

But it would be a mistake to see the 1980s as the final chapter of the Chicago School. Four major movements in Chicago economics since 1980 are captured in the awarding of more recent Nobel Prizes. Gary Becker was awarded the prize

for his work in the new home and social economics. Robert Lucas won for developments in empirical macroeconomics. Merton Miller was joined by former Chicago researcher Harry Markowitz for their development of finance theory. And James Heckman won the prize for the development of microeconometrics. Alongside these scholars (Miller died in 2000, but the others remain active), the next generation of Chicago economists is making a place for itself. Both Thomas Sargent and Lee Hansen have won the new Erwin Plein Nemmers Prize for significant contributions to new modes of analysis in economics, and Kevin Murphy and Steven Levitt have won the coveted John Bates Clark medal from the American Economic Association.

Each of the four recent movements within Chicago economics – finance, empirical macroeconomics, the new home economics, and microeconometrics – are rooted in common Chicago themes: the application of price theory, the development of methods for the quantitative analysis of social problems, and the notion that economics is an applied policy science. The Chicago approach rests on a three-legged stool which combines an appreciation for the 'simple' analytics of Marshallian price theory (as Reder observes, a constant at Chicago since the early 1930s), the development of quantitative tools as expressed in Friedman's classic article (1953) on 'positive economics,' and the Becker–Stigler prescription to focus attention on the elements of the constraint set, rather than changes in values and preferences, in the explanation of human behaviour (see Becker 1976; Stigler and Becker 1977). Once combined, this three-legged methodological stool provided a stable foundation for the continued expansion of the scope of social scientific problems that Chicago economists have addressed (Becker 1981; Becker and Murphy 2000; Levitt and Dubner 2005). Economic imperialism it may be, but Chicago economists argue that it is the only basis upon which a true social science can be built (see Lazear 2000).

Yet Reder's claim that the book on Chicago economics was about to close was right at least in one regard. Up to the mid-1970s, Chicago economists were an embattled minority (albeit growing in numbers and influence) of the economics profession. After the early 1980s, Chicago was no longer embattled, or even a minority. Its central ideas are still alive, but they are no longer the notions of a contrary-minded small group of scholars; in antitrust, law and economics, monetary theory, labour, finance and applied microeconomics, they comprise a position that has been widely adopted. Chicago economics today is part of the discipline's mainstream; indeed, in some sub-fields it has defined the mainstream. Success outside the confines of Chicago has also changed the School itself: since 1980, Chicago economics has gradually accommodated itself to the common standards of the discipline. Finally, the role of the Chicago School themes within the university has also been rendered more complicated by the remarkable expansion of the Graduate School of Business and the Law School as centres of Chicago-style economic, legal and public policy analysis.

Change and Continuity in Chicago Economics

The 1980s were a period of transition in Chicago economics, in several regards. For most of the period from the late 1940s until the early 1980s, the department of economics was chaired by either T. W. Schultz, D. Gale Johnson or Arnold Harberger; the required price theory course (ECON 301) was taught by either Milton Friedman, Harberger or Gary Becker, and required thorough familiarity with the canon of Chicago price theory – the theory texts of Knight (1933), Friedman (1962), Stigler (1966), Becker (1971), and Alchian and Allen (1969); and the other required first-year course was titled 'money' (not macroeconomics). The continuity in leadership was disrupted in the early 1980s (just as it had been 30 to 40 years earlier by the departures of Jacob Viner, Oskar Lange and the Cowles Commission, and the retirement of Frank Knight), as the early luminaries retired and passed responsibility on to the next generation (although Becker still shares some of the teaching in ECON 301). But a successful programme

is not built around individual scholars, even if they are luminaries like Friedman, Stigler, and Becker. Chicago's success, even in the period from the 1940s to the 1980s, is misunderstood if it is interpreted simply as the product of the unique cluster of scholars that it managed to attract (compare Van Overtveldt 2007, with Emmett 1998). In the early 1950s, the economics department replaced the traditional lone-scholar model of graduate education and faculty research with a workshop model that created an educational environment for graduate students and faculty members more closely akin to a scientific laboratory within which students and faculty pursued a collaborative intellectual project. While the Chicago model is reasonably well-known today and emulated, it was quite unique in the post-war period, and is central to Chicago's success. After passing the core examinations in price theory and money at the end of the first year, students not only continued to take courses but also associated themselves with a workshop (most workshops were open, so students often attended more than one; but each student was primarily associated with one workshop). Faculty were also associated with at least one workshop, and frequently defined the workshop's style: Friedman's money workshop; Stigler's industrial organization workshop; Fogel and McCloskey's economic history workshop; Harberger's Latin American finance workshop; and Coase's law and economics workshop. In the early years no common model had been established, and the workshops varied significant. Eventually, most workshops adopted the 'Chicago rules': the workshop met once per week, papers were distributed beforehand and therefore assumed to have been read, and presenters knew that discussion of the paper might begin as soon as five minutes into their presentation. Most of the workshop time was spent dissecting the paper's thesis, method, and data. Because the pattern of discussion was repeated every week in a dozen or more workshops, students and faculty became quite adept at working within Chicago's rules, applying Marshallian price theory to a wide range of policy-relevant topics. By the early 1980s, the number of economics workshops in the department, the Graduate School of Business, and the Law School was approaching 20. Today, in 2006, it still numbers in the teens.

The transition of key personnel in the early 1980s, therefore, did not affect the structure of the research and educational enterprise which supports the Chicago School. However, it did have an impact on the nature of the research and education of Chicago economists. By the end of the 1980s, the texts which comprised the canon of Chicago price theory lost their pride of place in the reading lists for ECON 301. At about the same time, the 'money' course (ECON 302) became a study of 'income, employment and the price level' built around standard Walrasian general equilibrium models that characterize macroeconomic analysis in most economics programmes. As well, the development of more sophisticated econometric models and techniques came to play a larger role in economic research at Chicago. 'Quantitative methods' was added as a core examination that all students had to pass in order to continue beyond the first year. In short, Chicago economics today looks a lot like economics everywhere else (in part, of course, because Chicago's approach is taught elsewhere and other programmes have created collaborative research environments like the Chicago workshops), although there remains a distinct Chicago 'flavour' that distinguishes it from MIT, Harvard, Berkeley and Yale, if not from Stanford, UCLA and Washington.

Change and Continuity in the Interpretation of Chicago Economics

Even as the contemporary evolution of Chicago economics continues to involve both continuity and change, our understanding of the history of Chicago economics has also evidenced both continuity and change. Reder's original essay was constructed on a model of Chicago economics which placed a small group of key individuals and their ideas at the centre of the School; one could envision his essay as an examination of concentric circles emanating out from the inner

circle that started with Viner and Knight and then included Friedman, Stigler and Becker. While not rejecting Reder's model entirely, historians have begun to construct a story of the development of Chicago economics that complicates the model significantly. Three aspects of Chicago School historiography can be highlighted to illustrate the direction of contemporary historical research on the School, and indicate the potential for further research. First, the transition from the Chicago economics of the inter-war period to the Chicago School of the 1950s and 1960s involved several significant changes. The elements of continuity that Reder emphasized remain – the pre-eminent role of price theory, for example – but discontinuities have crept in. Daniel Hammond's recent work on Milton Friedman's early career provides a glimpse into how that transition influenced even one of the mainstays of Chicago economics. Arguing against the continuity thesis about Chicago price theory articulated by Mirowski and Hands (1998), Hammond shows that Friedman had as much in common with NBER-style statistical work as he did with Knight's Chicago approach (Hammond 2005; see also Hammond 2008, and Rutherford 2008) In fact, even Friedman's famous methodological essay may be more a statement of his experiences with the NBER and the Statistical Research Group at Columbia University (associated with Harold Hotelling) than any earlier Chicago economist. In more recent work, Mirowski and van Horn (2008) argue that, whatever the continuities of Chicago's price theoretic tradition are, the Chicago School of the 1950s and 1960s was shaped more by new research projects initiated in the effort to define a new liberalism to in the Cold War period than it was by the classical liberalism of the Knight–Simons agenda in the 1930s and 1940s (see also Amadae 2003). Thus, while the Chicago School of Friedman and company should not be seen as a totally new tradition, historical reconstructions of their work have opened the door to further exploration of continuities and potential discontinuities between 'old' and 'new' Chicago.

We have already seen the second aspect of contemporary historical reconstruction in the earlier discussion of the institutional framework of the Chicago School. Rather than seeing individual scholars and their ideas transforming modern economics (as suggested even recently by Van Overtveldt 2007), contemporary historiography suggests that the intellectual success of the School was built upon a unique research infrastructure, focused in the workshops. Constructing the history of the workshops involves investigating the support network they developed, ranging from private foundation funding to international connections for research and students. Mirowski and van Horn (2008) focus on the role of the Volcker Fund, but other foundations and external research organizations like the Ford Foundation, Rockefeller Foundation (which funded many activities across the University of Chicago from its inception), Earhart Foundation, and the RAND Corporation participated in supporting Chicago's research infrastructure. In terms of international connections, much has been said of the role of the 'Chicago boys' in Chile, who set the groundwork for economic liberalization in Latin America and elsewhere, but were appointed to their positions by General Pinochet (Valdés 1995; Barber 1995). However, the institutional history of the Chile connection, which goes back to the early 1950s with an educational exchange between the University of Chicago and the Catholic University in Chile, has yet to be completely told. And we also do not have any histories of Chicago's other international research and student connections, including the equally unique relationship with the Hebrew University in Jerusalem and the University of Tel Aviv, despite the fact that Chicago was one of the few American academic institutions that welcomed Jewish scholars.

The third aspect of the Chicago School points toward two potential areas of research which would deepen the type of historical work illustrated above, while also providing insight into the degree of continuity and change within the School. Neither of these areas of research has made significant inroads into contemporary research. The first is the story of the integration of econometric developments at Chicago into the story of Chicago economics (as opposed to their place in the econometric literature). How did we go from Friedman and Stigler to Heckman,

Hansen, and Levitt? Was it just Chicago accommodating itself to the mainstream of the discipline, as is often suggested? Did Zvi Griliches and the development of quantitative analysis in agricultural economics play a role? Or Gregg Lewis and Albert Rees and labour economics? Did a quiet revolution go on at Chicago in the fields outside the core exams that gradually changed the School as a whole? These stories need to be examined in greater detail (see Kaufman 2008, for the history of Chicago labour economics in this regard). Second, the Chicago School's laissez-faire reputation is offset by the fact that a large portion of its graduates have gone into public service both in the United States and elsewhere. Harberger alone can count approximately 20 former students who have become central bank governors and ministers of finance. And countless Chicago students staff national and international economic ministries, commissions, and other organizations. If Chicago economists do believe that economics is a policy science, then the history of their interaction with policy, both as policy advocates and as policymakers, needs to be incorporated into our history. Again, what we do know about this history is piecemeal or quite general (for a start in the right direction; see Banzhaf 2008).

The new perspectives on Chicago economics open the door to both reconstructing the story of the Chicago School and to extending that story forward to the present. While Reder may have been premature to suggest the School's demise, both the reconstruction of its history and the story of its recent developments suggest both continuity and change.

See Also

- ▶ Becker, Gary S. (Born 1930)
- ▶ Buchanan, James M. (Born 1919)
- ▶ Chicago School
- ▶ Coase, Ronald Harry (Born 1910)
- ▶ Director, Aaron (1901–2004)
- ▶ Economic History
- ▶ Finance (New Developments)
- ▶ Friedman, Milton (1912–2006)
- ▶ Griliches, Zvi (1930–1999)
- ▶ Heckman, James (Born 1944)
- ▶ Human Capital
- ▶ Johnson, D. Gale (1917–2003)
- ▶ Johnson, Harry Gordon (1923–1977)
- ▶ Knight, Frank Hyneman (1885–1962)
- ▶ Labour Economics
- ▶ Laissez-Faire, Economists and
- ▶ Law, Economic Analysis of
- ▶ Lucas, Robert (Born 1937)
- ▶ Macroeconomics, Origins and History of
- ▶ Methodology of Economics
- ▶ Miller, Merton (1923–2000)
- ▶ Monetary Economics, History of
- ▶ New Classical Macroeconomics
- ▶ Patinkin, Don (1922–1955)
- ▶ Reid, Margaret Gilpin (1896–1991)
- ▶ Rosen, Sherwin (1938–2001)
- ▶ Scholes, Myron (born 1941)
- ▶ Schultz, Henry (1893–1938)
- ▶ Schultz, T. W. (1902–1998)
- ▶ Simons, Henry Calvert (1899–1946)
- ▶ Stigler, George Joseph (1911–1991)
- ▶ Viner, Jacob (1892–1970)
- ▶ Yntema, Theodore O. (1900–1985)

Bibliography

Alchian, A.A., and W.R. Allen. 1969. *Exchange and production: Theory in use*. Belmont, CA: Wadsworth.

Amadae, S. 2003. *Rationalizing capitalist democracy: The cold war origins of rational choice liberalism*. Chicago: University of Chicago Press.

Banzhaf, S. 2008. The Chicago school of welfare economics. In *The Elgar companion to the Chicago school*, ed. R.B. Emmett. Cheltenham, UK: Edward Elgar.

Barber, W.J. 1995. Chile con Chicago: A review essay. *Journal of Economic Literature* 33: 1941–1949.

Becker, G.S. 1971. *Economic theory*. New York: Knopf.

Becker, G.S. 1976. *The economic approach to human behavior*. Chicago: University of Chicago Press.

Becker, G.S. 1981. *A Treatise on the Family*. Cambridge, MA: Harvard University Press.

Becker, G.S., and K.M. Murphy. 2000. *Social economics: Market behavior in a social environment*. Cambridge: Belknap Press.

Emmett, R.B. 1998. Entrenching disciplinary competence: The role of general education and graduate study in Chicago economics. In *From interwar pluralism to postwar neoclassicism*, ed. M. Rutherford and M. Morgan. Durham, NC: Duke University Press.

Friedman, M. 1953. The methodology of positive economics. In *Essays on positive economics*. Chicago: University of Chicago Press.
Friedman, M. 1962. *Price theory: A provisional text*. Chicago: Aldine.
Hammond, D.J. 2008. The development of postwar Chicago price theory. In *The Elgar companion to the Chicago school of economics*, ed. R.B. Emmett. Cheltenham, UK: Edward Elgar.
Hammond, D.J. 2005. More fiber than thread? Evidence on the Mirowski–Hands yarn. In *Agreement on demand: Consumer choice theory in the 20th century, annual supplement to history of political economy*, ed. P.E. Mirowski and D.W. Hands. Durham, NC: Duke University Press.
Kaufman, B.E. 2008. Labor in Chicago economics. In *The Elgar companion to the Chicago school of economics*, ed. R.B. Emmett. Cheltenham, UK: Edward Elgar.
Knight, F.H. 1933. *The economic organization*. Chicago: University of Chicago Press.
Lazear, E. 2000. Economic imperialism. *Quarterly Journal of Economics* 115: 99–146.
Levitt, S.D., and S.J. Dubner. 2005. *Freakonomics: A rogue economist explores the hidden side of everything*. New York: William Morrow.
Mirowski, P.E., and D.W. Hands. 1998. A paradox of budgets: The postwar stabilization of neoclassical demand theory. In *From interwar pluralism to postwar neoclassicism, annual supplement to history of political economy*, ed. M.S. Morgan and M. Rutherford. Durham, NC: Duke University Press.
Mirowski, P.E., and R. van Horn. 2008. Neoliberalism and Chicago. In *The Elgar companion to the Chicago school of economics*, ed. R.B. Emmett. Cheltenham: Edward Elgar.
Rosen, S. 1974. Hedonic prices and implicit markets: Product differentiation in perfect competition. *Journal of Political Economy* 82: 34–55.
Rosen, S. 1981. The economics of superstars. *American Economic Review* 71: 845–858.
Rosen, S. 2004. *Markets and diversity*. Cambridge, MA: Harvard University Press.
Rutherford, M. 2008. Chicago economics and institutionalism. In *The Elgar companion to the Chicago school of economics*, ed. R.B. Emmett. Cheltenham, UK: Edward Elgar.
Stigler, G.J. 1966. *The theory of price*, 3rd ed. New York: Macmillan.
Stigler, G.J., and G.S. Becker. 1977. De gustibus non est disputandum. *American Economic Review* 67: 76–90.
Valdés, J.G. 1995. *Pinochet's economists: The Chicago school in Chile*. Cambridge: Cambridge University Press.
Van Overtveldt, J. 2007. *The Chicago school: How the University of Chicago assembled the thinkers who revolutionized economics and business*. Chicago: Agate.

Child Health and Mortality

Janet Currie

C

Abstract

Child health is a major indicator of the direction and well-being of society. It is a significant factor predicting health and productivity in adult life, and the health of adults in turn affects the well-being of the next generation of children. The most important outstanding issues include determining the most cost-effective investments in child health, explaining the relationship between health and socioeconomic status over the life course, and finding the interventions that are most effective in breaking the inter-generational cycle of ill health and poverty. As children are economic actors in their own right, their well-being is worthy of study.

Keywords

Anthropometrics; Asymmetric information; Child health and mortality; Education; Family economics; Fertility in developed countries; Fertility in developing countries; Health care; Health economics; Health insurance; Human capital; Mortality; Pollution; Poverty alleviation programmes; Productivity; Public health; Risk

JEL Classifications

J10

Child health and mortality are of interest to economists for three reasons. First, they are important indicators of the success or failure of government policy. Second, children's health has long-term impacts on their health and productivity as adults. Third, there is increasing recognition that children are economic actors in their own right. Hence, their well-being is worthy of study.

The most common model of child health is one in which health is 'produced' by families using

health 'inputs' (Grossman 2000). Examples of inputs include the goods and services families buy to improve child health. Families maximize an intertemporal utility function subject to the production function, prices, and budget constraints. Inputs are valued only because of their effect on health. Children start with a 'health endowment' that depreciates over time in the absence of health inputs. Public policy affects either the price of inputs or the form of the production function. The model predicts that child health will be influenced by the price of health inputs. The inter-temporal nature of the model highlights the idea that health inputs are investments with long-term payoffs.

Studies of children in developing countries often focus on the 'production' of mortality rates, nutrient intakes, height, weight and other objective measures. In contrast, studies of children in richer countries often focus on the utilization of medical care. But health care is only one input into the production of child health, and it is not the most important. Improvements in standards of living, advances in knowledge about disease and hygiene, and public health measures such as improved sanitation have done more to improve child health in the past 150 years than even the most spectacular advances in personal medical care (Preston 1977). Today, accidents and violence, rather than disease, are the major killers of young children in wealthy countries after the first year of life (UNICEF 2001).

Measures of Child Health

Health is multidimensional and difficult to measure. Mortality and parent-reported health fall at two ends of a spectrum. Mortality is an objective but narrow measure. In countries with high death rates, child mortality is a relatively sensitive indicator of economic and social conditions. For example, in Zimbabwe mortality among children under five years old increased from 80 to 126 per 1,000 live births between 1990 and 2003 as the economic crisis deepened (United Nations Common Database). In countries with lower child mortality rates, the relationship between economic conditions and mortality may be masked by the effects of economic cycles on fertility. For example, some recent papers demonstrate that in developed countries poorer people have fewer children during economic downturns so that the average health of infants increases (see, for example, Lleras-Muney and Dehejia 2004). The relationship between mortality and economic conditions is also masked by strong underlying downward trends in mortality due to technological advances.

A typical survey question eliciting parent reports about child health asks respondents to rate child health on a scale of 1 to 5. An advantage of this measure is that it applies to all children. A disadvantage is that parent reports may be biased. For example, sick parents are more likely to report sick children. Parents are also often asked about limitations on children's activities (for example: Did a health problem prevent school attendance?) and about the presence of chronic conditions. These questions have the advantage of being more specific, but capture only one dimension of health and also suffer from potential biases (Baker et al. 2004; Strauss and Thomas 1996).

In between are anthropometric measures such as birthweight, height, weight, height for age, and body mass index (Martorell and Habicht 1986). Anthropometrics are objective measures that apply to large numbers of children. But, like mortality, they may not be sensitive measures in healthy populations. For example, American children are unlikely to be stunted (low height for age) and are increasingly likely to survive low birthweight (less than 2,500 grams) without significant impairments. American children are increasingly likely to be obese, however, suggesting that body mass index is likely to become a more important health indicator in the future.

A fourth class of measures involve 'risky behaviours' such as precocious or dangerous sexual activity, involvement in crime or victimization, use of handguns, and use of alcohol, tobacco, and illegal drugs. Given the importance of accidents and violence among children, these are important questions. But the stigma associated

with these activities makes it likely that they will be under-reported. Also, risky behaviours may or may not lead to poorer health. Unfortunately, the actual health effects of many behaviours are very poorly reported. For example, there is little information available about injuries that do not lead to deaths.

Some surveys include clinical assessments of children's health by doctors or other trained professionals in addition to some of the information about economic status that is usually collected in social surveys. Examples include the British birth cohort studies, the American National Health and Nutrition Examination Surveys, the World Bank's Living Standards and Measurement Surveys and the Indonesia Family Life Survey. Some of the most interesting work being done in this area involves measures of children's genetic make-up. Caspi et al. (2002) show, for example, that New Zealand men with a specific genetic marker were more likely to be violent adults, but only if they had been maltreated as children.

Given the broad range of health outcomes, researchers should look at a range of outcomes and carefully consider whether the chosen ones are likely to be affected by the phenomena under study.

Health Care Utilization

The human capital model makes a clear distinction between health and health inputs. In the model, parents care about health rather than health inputs. Yet this distinction is often blurred. Williams and Miller (1992, p. 991) state that 'One of the most impressive aspects of health policy implementation [in Europe is] that the programs were put in place not because of extensive documentation on cost effectiveness, but out of a value system that cherishes equity in health care.' The underlying assumption is that all health care produces health. Yet the market for health care is plagued with imperfections. Some care is likely to be superfluous, for consumption rather than investment purposes, or even injurious.

Models of physician-induced demand show that asymmetric information can lead to excessive consumption of medical services if physicians take advantage of their superior information to 'sell' services that patients do not need (Pauly 1980; Dranove 1988). There may be considerable scope for inducement in the market for children's health care. Many child treatments are inexpensive but have a high clinical value when they are warranted, so parents perceive a low cost set against a potentially high benefit. The availability of insurance compounds the problem by further reducing costs to parents.

Researchers should focus on measures of utilization that have a clear benefit. Whether or not a child visited a doctor in a year and whether a child is immunized are good examples. Measures such as the number of hospitalizations are problematic since many hospitalizations could be prevented with appropriate outpatient care. Some recent work focuses on 'preventable hospitalizations' as a measure of inadequate utilization of care (Casanova and Starfield 1995).

Health as an Investment

Child health affects adult health. Poor health in childhood also lowers future utility through its effects on future wages and labour force participation (Currie and Madrian 1999) and through its effects on schooling. Currie (2005) provides a survey of literature linking several specific health conditions to cognitive outcomes and schooling achievement.

Using data from the 1999 Panel Study of Income Dynamics, James Smith (2005) shows that a retrospective self-reported question about health during childhood is remarkably predictive of future outcomes. Comparing siblings, he finds that those who were in excellent or very good health earn 25 per cent more as adults. Currie (2000) surveys some of the many studies that find positive associations between cognitive test scores and anthropometric measures of health such as birthweight, weight, height, head circumference, and the absence of abnormalities in children of various ages. More recently Currie and Moretti (2005) have shown that differences in birthweight between sisters are predictive of differences in education and median income in the

zip code of residence at the time the sisters deliver their own children many years later.

But low birthweight is only one of a number of health shocks that low-income children are more likely to experience (Newacheck et al. 1996). Case et al. (2002) show that the gap in health status between rich and poor US children widens as children age. Currie and Stabile (2003) replicate this finding using Canadian data, and argue that the widening gap reflects the greater frequency of negative health shocks among poor children. The comparison between the United States and Canada suggests that public health insurance is not sufficient to shield children from the negative health consequences of poverty (since Canada has universal insurance). However, in Britain the gap between rich and poor children is smaller than in North America and does not widen as children age (Currie et al. 2004). This suggests that some other aspect of the social safety net may be responsible for protecting child health in Britain.

Poor children are more likely than rich children to suffer from mental health problems (Currie 2005, 2002). Mental health problems account for the largest share of days lost due to health problems in the United States. Many mental health conditions have their roots in childhood, but the relationship between mental health and child outcomes has been largely ignored in economics. Currie and Stabile (2005) investigate the relationship between symptoms of Attention Deficit Hyperactivity Disorder Activity disorder (ADHD) and educational attainment using US and Canadian panel data. We find large negative effects even in rich sibling-fixed effects models. Other research has shown that childhood behaviour problems predict negative future outcomes (cf. Gregg and Machin 1998). The prevalence and potential economic importance of child mental health problems suggest that more work is warranted.

Policy and Child Health

It is easy to justify government intervention in the market for health care. In addition to asymmetric information between patients and providers, there are other informational problems. For example, imperfect information in the market for insurance can lead to market failure. And although parents make most decisions about child health inputs, these decisions have consequences for society. Parents who do not take account of externalities may not provide the optimal level of care for their children (cf. Kremer and Miguel 2004). Finally, the health sector accounts for a large and growing share of the economy, and the government is already the major player in the health care markets in most countries, including the United States. Policies can be divided into those that intervene in the market for health care and those that affect health through other means. Public health insurance is the most prominent example in the first category. It is difficult to study the impact of universal health insurance because there is only a single 'before/after' comparison. But over the late 1980s and early 1990s, the United States greatly expanded its public health insurance coverage of pregnant women and children. Forty per cent of US births are now covered by public insurance. The expansion took place at an uneven rate across states, yielding a potential source of identification.

The effects of this expansion of insurance coverage are surveyed in Gruber (2003). It reduced infant and child mortality, increased utilization of preventive care, and reduced preventable hospitalizations among children. But increases in coverage also increased the inappropriate use of care (for example, increased rates of Caesarean section). And some who took up public health insurance would have had private health insurance in the absence of the expansions. Hence, public health insurance improves child health, but does not necessarily result in efficient service delivery.

Health care utilization is only one input into health production. Other inputs such as a healthy lifestyle and the avoidance of injury are arguably much more important. Government policy has a large role to play in affecting many health inputs beyond health care. A few examples follow.

Pollution is likely to be more harmful to children than to adults both because they are still developing and because of their small size.

Hence, any policy that affects the environment may affect child health. For example, Chay and Greenstone (2003) show that the recession of the early 1980s reduced infant mortality. Currie and Neidell (2005) show that reductions in carbon monoxide pollution in California over the 1990s (largely due to cleaner vehicles) saved at least 1,000 infant lives.

Child obesity is a growing problem that threatens future health. The potential role for government ranges from the provision of information (for example, revising the 'healthy food pyramid' to reflect the most recent nutritional knowledge) to regulation (for example, eliminating Coke machines in schools). The government plays a similar role with respect to discouraging children from using alcohol and tobacco, though in these examples government also directly controls the price of the products through taxation. A good deal of research documents the relationships between prices, advertising, and youth consumption of tobacco and alcohol. But we know much less about the effectiveness of newer policies aimed at curbing obesity (see Gruber 2001).

Although injuries remain a major cause of death, the incidence of accidental death has declined dramatically since the 1970s, especially in the United States (UNICEF 2001). Glied (2001) argues that the decline is due to improvements in education resulting in increased use of, for example, bicycle helmets and seat belts. But many products, including cars, cribs, and medicine bottles, are much safer than they used to be. Is this a result of random technical innovation, government mandates, or fear of lawsuits? Similarly, trauma care has improved greatly. So there are many possible explanations for the reduction in mortality.

While health affects education, maternal education affects child health. Currie and Moretti (2003) find that increases in the availability of colleges increased women's education, leading to better infant health outcomes. Hence, there is an intergenerational payoff to government investments in education that leads to 'increasing returns' to investments in education (Rosenzweig and Wolpin 1994). Finally, as discussed above, poor children are more likely than rich children to suffer virtually all forms of health insult. Hence, improving health is a goal of general poverty alleviation programmes such as public housing and income maintenance.

Summary

Child health is an important indicator of the direction and well-being of society. Health in childhood is one of the more important factors predicting health and productivity in adult life, and the health of adults will in turn affect the well-being of the next generation of children.

Many policies have impacts on child health. Some simple improvements in data collection efforts could have a large research payoff in terms of identifying these impacts. These include: allowing the release of geographical identifiers so that health data can be merged to other data; the inclusion of family income and demographics in health data-sets; and the collection of more objective measures of child health. What are the most interesting outstanding questions? First, what are the most cost-effective investments in child health? Second, what explains the relationship between health and socio-economic status over the life course? And third, what interventions are most effective in breaking the inter-generational cycle of ill health and poverty?

See Also

- Family economics
- Fertility in developed countries
- Fertility in developing countries
- Health economics
- Household production and public goods
- Human capital

Bibliography

Baker, M., C. Deri, and M. Stabile. 2004. What do self-reported, objective measures of health measure? *Journal of Human Resources* 39: 1067–1093.

Casanova, C., and B. Starfield. 1995. Hospitalizations of children and access to primary care: A cross-national

comparison. *International Journal of Health Services* 25: 283–294.
Case, A., D. Lubotsky, and C. Paxson. 2002. Economic status and health in childhood: The origins of the gradient. *American Economic Review* 92: 1308–1334.
Caspi, A., et al. 2002. Role of genotype in the cycle of violence in maltreated children. *Science* 297: 851–854.
Chay, K., and M. Greenstone. 2003. The impact of airpollution on infant mortality: Evidence from geographic variation in pollution shocks induced by a recession. *Quarterly Journal of Economics* 118: 1121–1167.
Currie, J. 2002. Child health in developed countries. In *Handbook of health economics*, ed. J. Newhouse and A. Culyer. Amsterdam: North-Holland.
Currie, J. 2005. Health disparities and gaps in school readiness. *The Future of Children* 15(1): 117–38 (issue on *School Readiness: Closing Racial and Ethnic Gaps*).
Currie, J., and B. Madrian. 1999. Health, health insurance and the labor market. In *Handbook of labor economics*, ed. D. Card and O. Ashenfelter. New York: North-Holland.
Currie, J., and E. Moretti. 2003. Mother's education and the intergenerational transmission of human capital: Evidence from college openings. *Quarterly Journal of Economics* 118: 1495–1532.
Currie, J., and E. Moretti. 2005. *Biology as destiny: short and long-run determinants of intergenerational correlations in birth weight*, Working paper No. 11567. Cambridge, MA: NBER.
Currie, J., and M. Neidell. 2005. Air pollution and infant health: What can we learn from California's recent experience? *Quarterly Journal of Economics* 120: 1003–1030.
Currie, J., and M. Stabile. 2003. Socioeconomic status and health: Why is the relationship stronger for older children? *American Economic Review* 93: 1813–1823.
Currie, J., and M. Stabile. 2005. *Child mental health and human capital accumulation: The case of ADHD*, Working Paper No. 10435. Cambridge, MA: NBER, 2004, updated August 2005.
Currie, A., M. Shields, and S. Price. 2004. *Is the child health/family income gradient universal?* Evidence from England. Discussion Paper No. 1328. Bonn: IZA.
Dranove, D. 1988. Demand inducement and the physician/patient relationship. *Economic Inquiry* 26: 281–299.
Glied, S. 2001. The value of reductions in child injury mortality in the US. In *Medical care output and productivity*, ed. D. Cutler and E. Berndt. Chicago: University of Chicago Press.
Gregg, P., and S. Machin. 1998. *Child development and success or failure in the youth labour market*, Discussion Paper No. 0397. London: London School of Economics.
Grossman, M. 2000. The human capital model. In *Handbook of health economics*, vol. 1a, ed. A. Culyer and J. Newhouse. Amsterdam: North-Holland.
Gruber, J. 2001. *Risky behaviors among youths*. Chicago: University of Chicago Press.
Gruber, J. 2003. Medicaid. In *Means-tested transfer programs in the United States*, ed. R. Moffitt. Chicago: University of Chicago Press.
Kremer, M., and E. Miguel. 2004. Worms: Identifying impacts on education and health in the presence of externalities. *Econometrica* 72: 159–217.
Lleras-Muney, A., and R. Dehejia. 2004. Booms, busts, and baby's health. *Quarterly Journal of Economics* 119: 1091–1130.
Martorell, R., and J.-P. Habicht. 1986. Growth in early childhood in developing countries. In *Human growth: A comprehensive treatise*, ed. F. Faulkner and J. Tanner. New York: Plenum Press.
Newacheck, P., D. Hughes, and J. Stoddard. 1996. Children's access to primary care: Differences by race, income, and insurance status. *Pediatrics* 97(1): 26–32.
Pauly, M. 1980. *Doctors and their workshops*. Chicago: University of Chicago Press.
Preston, S. 1977. *Mortality trends. Annual Review of Sociology* 3: 163–178.
Rosenzweig, M., and K. Wolpin. 1994. Are there increasing returns to the intergenerational production of human capital? Maternal schooling and child intellectual development. *Journal of Human Resources* 29: 670–693.
Smith, J. 2005. *The impact of SES on Health over the life-course*. Santa Monica: RAND.
Strauss, J., and D. Thomas. 1996. Measurement and mismeasurement of social indicators. *American Economic Review* 86(2): 30–34.
UNICEF. 2001. *A league table of child deaths by injury in rich nations*. Innocenti Report Card, Issue No. 2. Florence: Innocenti Research Center.
United Nations Common Database. Online. Available at http://unstats.un.org/unsd/cdb/cdb_help/cdb_quick_start.asp. Accessed Mar 2005.
Williams, B., and A. Miller. 1992. Preventive health care for young children: Findings from a 10-country study and directions for United States policy. *Pediatrics* 89: S983–S998.

Child Labour

Kaushik Basu

Abstract

According to latest available estimates, somewhere between 14 to 18 per cent of all children between the ages of 5 and 14 years in the world

are labourers. The causes of child labour are many but the primary one is poverty, since for most parents sending children to work is an act of desperation. The availability of decent schools and the provision of small incentives, such as school meals, can help limit child labour. Hence, the best policy response is to improve conditions on the adult labour market, provide better schooling and, on rarer occasions, use legal interventions.

Keywords

Child labour; Education; Household production; Industrial Revolution; Poverty

JEL Classifications

O1

According to the International Labour Organization (ILO 2002) there were 186 million child labourers in the world in 2000, that is, children between the ages of 5 and 14 years doing regular economic work. This implies a 'participation rate' (the number of labouring children as percentage of all children of that age group) of 15.5 per cent. Of these, 111 million were engaged in 'hazardous work'. But by 2004 the number was down to 166 million – a participation rate of 13.7 per cent – and the number of children in hazardous work was down to 74 million. Some details and regional distribution estimates are available in Hagemann et al. (2006), but (at the time of writing) these new numbers are yet to be absorbed and analysed.

It is a truism that the incidence of child labour is hard to estimate, both because it is often illegal and so respondents would not proffer information too readily and because the work is usually in the informal sector where record keeping is weak. Not surprisingly, there are other estimates of child labour, higher and lower. According to the UNICEF (2006), which collates data from different sources from 1998 to 2004, the participation rate is 18 per cent.

These data sources have both upward and downward biases along different dimensions. Domestic work that is done in one's own household is usually recorded very poorly or not at all. But we have micro evidence that in poor regions children, especially girls, do huge amounts of work in their homes, ranging from fetching wood to hazardous work like cooking over open fires. Indirect evidence for this comes from the gender breakdown of child labour. According to ILO data, boys do more labour than girls; their participation rates are respectively 15.9 per cent and 15.2 per cent. But detailed micro studies that try to include heavy domestic work, such as that by Cigno and Rosati (2005, ch. 5), show that girls tend to do 30 per cent more work than boys. Hence, there is a downward bias in the macro numbers mentioned above.

On the other hand, one source of upward bias comes from 'work' being equated with doing more than one hour of work in the 'reference period', and from the fact that for most studies the reference period is one week. It is arguable that children who answer 'yes' because they barely satisfy that cut-off ought not to be classified as child labourers.

The reason for not becoming too weighed down by these statistical debates is that, no matter how one measures it and, as a consequence, whether the participation rate turns out to be 14 per cent or 18 per cent, it is easy to agree that the incidence of child labour is unacceptably high. In a world with as much opulence as ours there should not be so many children working and that too in grinding poverty and in intolerable working conditions.

This raises the question of the causes of child labour and the appropriate policy response. The primary cause is poverty. Well-off parents living in the same nation and under the same laws as poor ones almost never send their children to work. Hence, a child's non-work (whether this be leisure or schooling) is a luxury good. Sufficiently poor parents cannot afford this. This was called the 'luxury axiom' in Basu and Van (1998), and there is ample empirical evidence for it (see discussion in Ray 2000; Basu and Tzannatos 2003; Edmonds and Pavnick 2005). But there are other causes as well. There are parents on the borderline of poverty, who, if they knew that there were decent schools in the area and/or that their

children would get a square meal in school, would take the children out of labour and send them to school. Hence, the provision of schooling and, ideally, having some added incentives for sending children to school can make a large difference to the incidence of child labour (Ravallion and Wodon 2000; Bourguignon et al. 2003).

The presence of other determinants is also evident from the fact that the location of a child in the rural–urban spectrum affects the probability of the kind and amount of work the child is likely to do. This was always believed to be true. There were commentators at the time of the Industrial Revolution in Britain who argued that the alleged increase in child labour was really not an increase but a shift of child labour from agriculture to industry and a dramatic change in the *nature* of work (see Horrell and Humphries 1995, for discussion). Contemporary, casual evidence seems to support this. And a recent empirical study of child labour in Nepal (Fafchamps and Wahba 2006) formally confirms for the first time that urban proximity matters in a significant way. Children who live in or close to cities participate significantly less in labour and have a higher incidence of schooling than their rural counterparts. The health effects of these two kinds of child labour – agrarian and industrial – remain to be investigated systematically. Work in factories can be in dark and dank settings; on the other hand, agricultural work can mean exposure to not just the elements but also to pesticides and fertilizers. The net effects of these deserve investigation.

Given the multiplicity of causes, one has to be careful about the policy response to child labour. It is no surprise that, despite attempts by the British government from 1802 till the mid-19th century to deter child labour through a series of Factory Acts, the participation rate remained consistently and intolerably high. Indeed, the participation rates in Britain in the first half of the 19th century were higher than those found in today's China or India. Likewise in the USA, despite a variety of legislative measures starting in 1837 in Massachusetts, the incidence of child labour remained high and in fact continued to rise till the end of the 19th century.

While there is no final word on policy, we know that some measures are likely to be more effective than others. Ameliorating poverty, improving adult labour market conditions and providing better schooling, as already discussed, can have a significant effect. The law – bans and fines – can also play a role but should be used with caution and after empirical tests of whether the context deserves such measures. It has been argued (see Basu and Van 1998; Dessy and Pallage 2001; Emerson and Souza 2003) that the labour market can in different ways (such as the general equilibrium impact on market wages, coordination with technology and intergenerational dynamics) give rise to multiple equilibria. That is, the market, left to itself, can settle into different grooves; for instance, one with no child labour and another with a high participation rate. In such a case, if the economy settles into the latter equilibrium, a ban can be an effective tool. Otherwise a ban can lead children labouring in factories to worse outcomes, such as starvation or prostitution. Minimally, in such situations the law has to be combined with complementary interventions to ward off the extreme poverty and deprivation that can arise as a side effect of its implementation.

See Also

- ▶ Childcare
- ▶ Education in Developing countries
- ▶ Labour Economics
- ▶ Poverty

Bibliography

Basu, K., and Z. Tzannatos. 2003. The global child labor problem: What do we know and what can we do? *World Bank Economic Review* 17: 147–174.

Basu, K., and P. Van. 1998. The economics of child labor. *American Economic Review* 88: 412–427.

Bourguignon, F., F. Ferreira, and P. Leite. 2003. Conditional cash transfers, schooling and child labor. *World Bank Economic Review* 17: 229–254.

Cigno, A., and F. Rosati. 2005. *The economics of child labor*. Oxford: Oxford University Press.

Dessy, S., and S. Pallage. 2001. Child labor and coordination failures. *Journal of Development Economics* 65: 469–476.
Edmonds, E., and N. Pavnick. 2005. Child labor in the global economy. *Journal of Economic Perspectives* 19(1): 199–220.
Emerson, P., and A. Souza. 2003. Is there a child labor trap? *EconomicDevelopment and Cultural Change* 51: 375–398.
Fafchamps, M., and J. Wahba. 2006. Child labor, urban proximity, and householdcomposition. *Journal of Development Economics* 79: 374–397.
Hagemann, F., Y. Diallo, A. Etienne, and F. Mehran. 2006. *Child labour trends: 2000 to 2004*. Geneva: ILO.
Horrell, S., and J. Humphries. 1995. 'The exploitation of little children': Child labour and the family economy in the Industrial Revolution. *Explorations in Economic History* 32: 485–516.
ILO (International Labour Organization). 2002. *Every child counts: New global estimates on child labor*. Geneva: ILO.
Ravallion, M., and Q. Wodon. 2000. Does child labor displace schooling? *Economic Journal* 110: 158–175.
Ray, R. 2000. Child labor, child schooling and their interaction with adult labor. *World Bank Economic Review* 14: 347–367.
UNICEF. 2006. *State of the world's children 2006*. New York: UNICEF.

Child, Josiah (1630–1699)

Douglas Vickers

Keywords

Bullion; Child, J.; East India Company; Interest rates; Mercantilism; Transferable bills of exchange

JEL Classifications

B31

The second son of Richard Child, a London merchant, Sir Josiah Child was born in 1630 and enjoyed a highly successful merchant career during which he amassed a considerable fortune. His business ventures, which included the provisioning of Navy ships, led to his appointment as Deputy to the Navy's Treasurer at Portsmouth in 1655 and he became Mayor of that city in 1658. He was appointed a director of the East India Company in 1674, and with the exception of 1676 he was re-elected to a directorship in every subsequent year until his death. In 1681 he was elected governor of the company and established a close relationship with the Crown. Following the Revolution of 1688, and in response to mounting attacks on his conduct of company affairs, he relinquished some of his active management responsibilities.

Child's claim to recognition as an economist rests on his *Brief Observations concerning Trade and Interest of Money,* first published in 1668 and reissued (anonymously) in expanded form as *A Discourse about Trade* in 1690 and again as *A New Discourse of Trade* (with Child's name on the title page) in 1693. The work summarizes the views he presented to the Council of Trade appointed by the King in 1668 (following the appointment of a Select Committee on the State of Trade by the House of Commons in the preceding year) and to a similar House of Lords Committee in 1669.

Among the reasons for the mercantile supremacy of the Dutch, he cites the establishment of banks and the widespread use of transferable bills of exchange, which he strongly argued should be adopted in England. He argued for a reduction of the legal maximum rate of interest from six to four per cent (referring to this as 'my old theme'), claiming that the lower rate of interest in the Netherlands was 'the *causa causans* of all the other causes of the riches of that people'. He saw the beneficial effects on trade of a lower cost of money capital, but he did not discuss, as did John Locke at the same time, the relation between a legally established rate of interest and the rate established by natural market forces.

Child's argument that the beneficial effect of lower interest rates would cause 'all sorts of labouring people that depend on trade (to be) more constantly and fully employed' took up the then widespread concern with the employment problem and he concluded: 'it is our duty to God and nature so to provide for and employ the poor'. A significant discussion of the question of the poor and a scheme for their relief and

employment is included in Chapter II of the *Discourse of Trade.*

Notwithstanding his scattered observations that appear to support free trade principles and his assertion of the principles of competitive markets, Child was an exponent of monopoly when it suited his and the East India Company's advantage. He recognized the need to export bullion if that gave rise to further export trade opportunities. But his work abounds in arguments for trade restrictions in specific cases, such as those requiring the transportation of traded commodities in English vessels and requiring that colonial trade should be conducted only with England, thereby emphasizing the domestic employment-creating effects of the colonies. He stands as a latter-day mercantilist rather than an analytical anticipator of the laissez-faire doctrines of genuine and generalized freedom of trade.

Selected Works

1668. *Brief observations concerning trade and interest of money.* London.
1693. *A new discourse of trade*. London.

Bibliography

Letwin, W. 1959. *Sir Josiah child, merchant economist*. Boston: Harvard Graduate School of Business.
Macauley, T. (Lord) 1848, 1855, 1861. *The history of England from the accession to James the second*. New York: AMS Press, 1968.
Schumpeter, J.A. 1954. *History of economic analysis*. New York: Oxford University Press.

Childcare

David M. Blau

Abstract

The market for childcare and the role of the government in the childcare market have grown enormously as mothers of young children have entered the labour force in very large numbers. Economists have made important contributions to understanding many aspects of childcare. This article focuses on (*a*) the effect of the price of childcare on labour force participation of mothers of young children, (*b*) the effect of childcare and early childhood interventions on children, and (*c*) the rationale for and effects of government involvement in childcare. Fruitful avenues of additional research are suggested.

Keywords

Adverse selection; Child development; Childcare; Childcare subsidies; Education production function; Head start; Human capital; Imperfect information; Labour force participation; Market imperfections; Moral hazard; Poverty; Random assignment; Selfsufficiency; Women

JEL Classifications

J13

The market for childcare in advanced economies has grown enormously in response to the dramatic increase in labour force participation by mothers of young children. In 1950 12 per cent of married women in the United States with children under age six were in the labour force, compared to 63 per cent in 2000. Labour force participation of single mothers with children under six has also increased rapidly, reaching 65 per cent in 2000. As the market has grown, the role of the public sector in subsidizing, regulating, and providing childcare has increased substantially. One- third of all expenditure on childcare and preschool in the United States is financed by government subsidies or by direct provision of services. The public sector plays an even larger role in childcare in many European countries. Three aspects of childcare have received the most attention from economists: (*a*) the effect of the price of childcare on labour force participation of mothers, (*b*) the effect of childcare and early childhood interventions on child development, and (*c*) the rationale for and effects of government involvement in

childcare. Childcare is interpreted broadly here to include care provided by someone other than a child's parent either to facilitate employment of parents or to enhance child development.

Blau and Currie (2006) summarize the findings of 20 studies that estimate the elasticity of maternal labour force participation with respect to the price of childcare. The estimates vary widely across studies, but studies that account for the availability of informal unpaid childcare options usually estimate relatively small elasticities, in the range of − .09 to − .20. These studies use a multinomial choice framework that allows for the possibility that a mother can work without using paid childcare. Use of unpaid childcare by family members, relatives, and others is very common. The relatively small elasticity estimates suggest that a price increase induces substitution of informal unpaid childcare for paid care, dampening the sensitivity of maternal employment to the price of childcare. Some evidence suggests that the price elasticity is larger in absolute value for lower-wage women. This evidence confirms that childcare costs are a significant but not major barrier to employment of mothers. The evidence also implies that childcare subsidies increase work incentives of mothers, a finding confirmed by a small number of studies that directly analyse the impact of subsidy programmes on employment.

An important concern about childcare is that low-quality care could be harmful to the development of young children. Conversely, high-quality care may help compensate children from low-income families for the disadvantages of growing up in poverty. The effect of childcare on child development has traditionally been the domain of developmental psychology, but in recent years economists have contributed to this literature, noting its similarities to the 'education production function' literature for school-age children.

The quality of childcare can be characterized by 'structural' features such as the size of the group in which care is provided, the ratio of adult caregivers to children, and the education and specialized training of providers. Alternatively, direct observation of the developmental appropriateness of the care received by children can be made by trained observers using standardized instruments. These 'process' measures of quality are more proximate determinants of child development than are the structural features.

The small amount of evidence available suggests that higher-income parents do not choose higher-quality childcare on average: among users of day-care centres, there is no systematic relationship between family income and the quality of childcare used, if other factors are controlled for (Blau 2001). This is true whether the quality of care is measured by structural characteristics or process measures. This suggests that parents are either unable to discern the quality of care, or unwilling to pay the additional cost associated with higher-quality care, or both.

Several random assignment demonstration projects have evaluated the impact of high-quality preschool programmes for disadvantaged children (see reviews in Blau 2001; and Blau and Currie 2006). The results show that such programmes have delivered substantial long-run benefits to the participants and society: lower school dropout rates, higher earnings, fewer out-of-wedlock births, and lower public expenditures on welfare, criminal justice, and special education. Benefit-cost calculations show that these interventions have a very high social rate of return. This evidence is compelling, but it is based on very intensive and costly programmes that are of exceptionally high quality and are targeted at highly disadvantaged children. It is unclear whether childcare of moderately high quality provides positive but proportionately smaller developmental benefits, or whether there exists a threshold of quality below which benefits are negligible. It is also unclear how the quality of childcare affects children who are not highly disadvantaged. In non-experimental studies that follow children over time, higher-quality childcare is associated with better developmental outcomes in the short run (one to three years). However, it remains uncertain to what extent this is a causal impact. Recent studies that control for many other potentially confounding factors find that the quality-development association is smaller than in models with fewer controls, but remains significantly different from zero.

Two main arguments have been used to rationalize a role for government in the childcare

market. The arguments are based on attaining economic self-sufficiency, and childcare market imperfections. On self-sufficiency, childcare subsidies might help low-income families achieve economic self-sufficiency, defined as being employed and not enrolled in welfare programmes. Self-sufficiency is a desirable goal because it may inculcate a work ethic and generate human capital through on- the-job training and experience. These arguments explain why many childcare subsidies require employment or work-related activities such as education and training. Subsidies for childcare and other work-related expenses paid to employed low-income parents may cost the government more today than would cash assistance. But these subsidies could result in increased future wages and hours worked and lower lifetime government support than the alternative of cash assistance both today and in the future. This argument has nothing to do with the effects of childcare on children, and there are few restrictions on the quality of childcare that can be purchased with employment-related childcare subsidies. However, evidence on wage growth of low-skill workers suggests that wages grow only modestly with experience, too slowly to lift low-skill workers out of poverty (Gladden and Taber 2000). Middle and upper-income families are generally not at risk of going on welfare, so it is not obvious that there is an economic rationale for subsidies for their employment-related childcare expenses.

As for market imperfections, the imperfections that are often cited are imperfect information available to parents about the quality of childcare, and positive external benefits to society generated by high-quality childcare (Walker 1991). Imperfect information exists because consumers do not know the identity of all potential suppliers, and the quality of care offered by any particular supplier is not fully known. A potential remedy for the first problem is government subsidies to resource and referral agencies to maintain comprehensive and accurate lists of suppliers. The second information problem arises because consumers know less about product quality than does the provider, and monitoring the provider is costly to the consumer. This can lead to moral hazard and/or adverse selection. The limited evidence available suggests that parents are not well-informed about the quality of care in the arrangements used by their children. Childcare subsidies targeted at high-quality providers could induce parents to use higher-quality care.

The externality argument is a standard one that closely parallels the reasoning applied to education. High-quality childcare leads to improved intellectual and social development, which in turn increases school readiness and completion. This reduces the cost to society of problems associated with low education: low earnings, unstable employment, crime, drugs, teenage childbearing, and so forth. If parents are not fully aware of these benefits, or account for only the private and not the social benefits, then they may choose childcare of less than socially optimal quality. This argument could rationalize subsidies targeted to high-quality providers, such as Head Start, a US programme aimed at enhancing cognitive and social development of low-income children.

As this discussion implies, childcare policy can be used to facilitate employment of mothers and enhance development of young children. There is likely to be a trade-off between these goals because higher-quality care is more expensive. There is not a political agreement in the United States to spend enough to achieve both goals, or on which goal should have the highest priority. This is due in part to conflicting views on the proper role of the government in a domain that was mainly left to families until the last quarter of the twentieth century. But it also reflects lack of knowledge about the magnitudes of important parameters that affect the costs and benefits of alternative policies. Economists could make significant contributions to knowledge by careful empirical studies that produce reliable estimates of such parameters. The following issues seem important and well-suited to analysis by economists. Despite a large number of studies, there is considerable uncertainty about the magnitude of the elasticity of maternal employment with respect to the price of childcare. A careful sensitivity analysis could help resolve this uncertainty. Research on the price-responsiveness of low-income mothers would be especially useful.

Consumer demand for quality in childcare is not well-understood, and new research could be valuable. Research on the take-up decisions of families eligible for childcare subsidies would be useful in order to determine the likely effectiveness of different forms of subsidies. New research on the supply of childcare would be useful. Subsidies to consumers may bid up the price of childcare, and it is important to be able to quantify such effects. It would also be useful to examine the quality supply decisions of providers, in order to determine how responsive the supply of high-quality care might be to subsidies.

See Also

- ▶ Education Production Functions
- ▶ Family Economics
- ▶ Labour Supply
- ▶ Women's Work and Wages

Bibliography

Blau, D. 2001. *The child care problem: An economic analysis*. New York: Russell Sage Foundation.

Blau, D., and J. Currie. 2006. Preschool, day care, and after school care: Who's minding the kids? In *Handbook on the economics of education*, ed. E. Hanushek and F. Welch. Amsterdam: North-Holland.

Gladden, T., and C. Taber. 2000. Wage progression among less skilled workers. In *Finding jobs: Work and welfare reform*, ed. R. Blank and D. Card. New York: Russell Sage Foundation.

Walker, J. 1991. Public policy and the supply of child care services. In *The economics of child care*, ed. D. Blau. New York: Russell Sage Foundation.

China, Economics in

Kiichiro Yagi

Abstract

Although pre-modern China possessed rich ideas pertaining to economic matters, they were not separated from the discourse of morality and politics. Even in the late 20th century, Chinese thinking, often unconsciously, reflected traditional ideas. Liberal economics missed the chance to guide the modernization of China. Marxian economics established its monopoly under the reign of the Communists. However, China had Marxian economists who supported its gradualist transition to a market economy. In the 1990s, the task of guiding economic reforms in China was handed over to a new generation of economists who absorbed 'Western' (non-Marxian) economics.

Keywords

Land nationalization; Asiatic mode of production; Central planning; China, economics in; Confucianism; Datong (Great Harmony); Equality; George, H.; Institutional economics; Law of value; Legalist School (China); Management buyouts; Marxism; Modernization; Mohist School (China); Political economy; Population growth; Public choice; Socialist commodity economy; Socialist market economy

JEL Classifications

B2

Economics in China has not been able to disassociate itself from politics. The Chinese word for economy or economics, *Jingji*, is the abbreviation of *Jingshi* (or *Jingguo*) *Jimin,* which means 'ruling the society or state and saving the people'. In traditional Chinese learning, this is a generalized concept that covers almost the entire range of a state's administrative activity. However, the viewpoint implied in this word is that of the rulers or administrators and not that of individuals engaging in economic activity on their own account.

The Quest for Wealth and the Control of Morality

Policies oriented towards the attainment of 'wealth and power' had appeared already in ancient China,

the Eastern Chou Period (722–256 BC), when the rule of Chou dynasty became in title alone and powerful vassal lords struggled with each other for leadership, which was based on the power of their feudal states. A crucial insight pertaining to economic growth that emerged during this period was that fostering the material welfare of the people was a precondition for a strong state. The famous saying 'Man will care about honour and disgrace only when he has enough clothing and food' is attributed to Guan Zhong (730–645 BC), the prime minister of a ducal state. He implemented policies that would bring stability to people's lives; these policies included the promotion of agriculture, monopolizing salt and iron, state intervention in the public distribution system, maintenance of a balanced budget and the consolidation of taxation and military services. Practical policies were further developed by many politicians in the Warring States Period (475–221 BC). These became part of the arsenal of policy measures adopted by the administrators of the unified state of successive dynasties from the Qin (221–206 BC) to the Qing (AD 1644–1911). A text named after Guan, *Guan Zi,* was compiled in the Western Han Period (206 BC–AD 8). This contains detailed discussions of the practical economic policies of ancient China.

In ancient China, before the unification by the Qin, political control over merchants was not strict. Wealthy merchants in the pre-Qin period were vividly described in *Records of the Historian* ('*Shiji*'). The editor–historian, Sima Qian (145–87 BC) clearly favoured a liberal economic policy that permitted the innovative activities of talented merchants.

Competitive Schools in Ancient China

Confucius (551–479 BC) also recognized the quest for wealth as a natural human trait. However, he stressed that the teachings of morality (*Ren*) should control the quest for wealth. According to him, superior men can understand and adhere to the virtues of righteousness and benevolence in their deeds, while inferior men (common people) cannot. The former belong to the ruling class and the latter are the ones who are ruled, who must be guided by the former. Confucius stressed the educational effect of a ruler on the people's perception of societal order. He was opposed to the levying of heavy taxes and unnecessary state intervention, since that might jeopardize the common man's standard of living. He maintained that a peaceful and fair reign of a virtuous ruler fosters allegiance. As long as people follow the basic order of society, the wealth of the state emerges as a spontaneous result of the growth in the population.

Meng Ke (c. 390–c. 305 BC), whose name is often mentioned together with Confucius, strictly excluded the consideration of material benefits from the political discourse of superior men. During his first meeting with the king of Liang, Meng declared that he spoke only of 'righteousness' (*Yi*) and not of 'benefits' (*Li*). However, he also stated that the dominance of 'righteousness' presupposes the maintenance of a 'permanent property' of the people in order to secure the morality of the people (*Mencius*).

Mo Di (c. 468–c. 367 BC) and his School (Mohists) grounded their altruistic teaching on the extended approval of 'benefits'. They believed that economic transactions are acts of 'mutual benefit', which will eventually support the doctrine of 'universal love'. From a utilitarian viewpoint, they regarded righteousness as a material benefit; this is in clear contradiction with the Confucians. Mohists further advocated a ban on war and simple burial. Apparently, this School originated from the craftsmen who were not entirely integrated into the social hierarchy existing in the pre-Qin period.

Legalists such as Shang Yang (c. 390–338 BC) and Han Fei (c. 310–238 BC) differed from the Confucians with respect to the measures to be adopted for guiding people. They stressed the effective control of people by the strict enforcement of punishment. They prioritized agricultural production and considered manufacturing and commerce as tertiary activities. The Legalists were prepared to collaborate with princes and politicians who sought to enhance the wealth and power of their states.

Omnipotent State Versus Virtuous Reign

Ancient China was unified by the Qin dynasty, which had adopted the policies of Legalists. The first emperor of the Qin (221–206 BC) suppressed Confucians who criticized his reign as measured against the criterion of virtuous ruler. However, under the following dynasty, the Han, Confucianism established its position as the state orthodoxy, which continued until the end of the Qing dynasty. Still, a Legalist direction survived in the pragmatic mentality and policies of administrative bureaucracy. Thus, Chinese political history witnessed repetitive conflicts between the moralistic direction of Confucianism and the bureaucratic administration in the direction of Legalists.

One of the most noteworthy debates was the dispute on salt and iron (81 BC), in which San Hongyoung (152–80 BC) – the finance minister of the Western Han dynasty – had to defend his policy against the criticism of Confucian scholars. In order to compensate for the deficit in the state finance caused by an expansionary policy, San extended the state monopoly of salt and iron and introduced a state- managed storage and distribution system. Such a system could be legitimized if it was successful in guaranteeing the nationwide provision of necessaries and a stabilization of their prices. However, coupled with a heavy tax burden, San's system made a devastating impact on the nation. Confucian scholars voiced the dissatisfaction of the people and pressed for the abolition of San's system.

A similar constellation appeared in the dispute around the economic reforms of Wang Anshi (1021–86). Wang's attempt to consolidate public finances by suppressing the annexation of lands by rich families and establishing a strict taxation system was opposed by traditional scholars, who were in alliance with the richer families.

Apart from the taxation and market control, Chinese administrators showed their expertise in the area of currency. They are the first to have issued paper money (*Jiao Zi*) in the 11th century. The Yuan dynasty (1271–1368) adopted the idea of inconvertible notes in its monetary system. The paper currency ordinance of 1287 drafted by Ye Li (1242–92) contained sound measures to maintain the value of paper money in relation to the regularly inspected silver reserve fund. This paper currency system of the Yuan dynasty exerted a certain influence over the currency system of other countries through the commercial networks under the grand Mongolian rule.

The Demand for Equalization

Support for equality is another persistent trait of traditional Chinese economic thought. The equalization of land and wealth was a typical demand raised by numerous peasant rebellions. The Taiping Rebellion (1851–64) put into effect an equal distribution of land, and the rural revolution under Mao Zedong's (1893–1976) directive displayed a similar kind of egalitarianism. However, the ideal of equality in the distribution of wealth can also be found in Confucian classics. Confucius himself remarked that rulers must worry 'not about the scantiness of wealth but its inequality of distribution' since 'there will be no feeling of poverty under equal distribution' (*Analects*). Here, equality is appreciated with respect to its ability to maintain harmony and tranquillity among the ruled. Meng Ke also proposed an egalitarian *Jin* land system, in which peasants, who were allotted equal amounts of land, jointly cultivated public land for the sake of generating public finance. This proposal was revived several times by reformist politicians as well as by egalitarian rebels.

A vision of an egalitarian ideal society, the Great Harmony (*Datong*), where neither private property nor egoistic interests exist, is mentioned in the Confucian classic *Li Ki. Xiaokang*, a society in which the people are guided by order and institutions is not an ideal but a second best, suited to the age of a civilized society. However, towards the end of the Qing dynasty, Kang Youwei (1858–1927), a reformist politician and scholar, revived the ideal of the Great Harmony to regenerate the whole nation.

Preconditions for Chinese Modernization

The nationwide examination system for the recruitment of government officials was established under the Sui dynasty (581–618) and continued until 1905. Based on the Confucian orthodoxy, it moulded the thought of Chinese intellectuals over a millennium. However, Confucian orthodoxy was not totally exempt from change. In addition to the ideas that had emerged in the ancient period, it absorbed heterogeneous ideas from other intellectual schools of thought, such as Buddhism and Taoism. The effect of the development of a rationalistic Neo-Confucianism guided by Zhu Xi (1130–1200) and the emergence of the countervailing school of Wang Yangmin (1472–1528), which introduced an inner integrity to Confucianism, are interesting issues that need to be further researched. Towards the end of the Ming dynasty (1368–1644), these developments promoted a critical attitude towards the traditional order of the empire. Huang Zongxi (1610–95) and Wang Fuzhi (1619–95) developed a utilitarian concept of hierarchy based on the private property and self-interest of the people. Further, the diffusion of the teaching of Wang Yangmin (*Xinxue*) that stressed purity of mind nourished the morality of the merchants (Yu 1987). However, these developments were not sufficient to modernize the Chinese intellectual tradition from within. The landlord class that recruited state officials through a nationwide examination formed the ruling alliance of the society. Merchants had no other option but to join this alliance as subordinate participants. However, the intellectual legacies of old China were preconditions for the Chinese to cope with the modernization that was initially forced on them by external forces.

Introduction of Western Economics

It was the publication by Wei Yuan (1794–1857) of the *Geography of the Maritime Countries* (1843) that initiated the movement among Chinese intellectuals of learning from the West. However, Western economics was not introduced until two decades later. Using H. Fawcett's *A Manual of Political Economy* as a textbook, W.A.P. Martin, an American Christian missionary, began a course on policies for the wealth of nations at a government school in Beijing in 1867. Later, in 1883, this course was translated and published in Shanghai under the same title. A second significant contribution pertaining to the translation of Western economics was that of a British missionary, J. Edkins, who translated W.S. Jevons's *Primer of Political Economy* into Chinese. This translation was published in 1886 with the Chinese title, *Policies for the Wealth of Nations and Support of People*. Fawcett and Jevons were neither mercantilists nor interventionists. However, both Chinese titles suggest that the Chinese people of this period regarded Western economics as a policy measure to strengthen the state.

Between 1901 and 1902, Yan Fu (1853–1921) published the translation of Adam Smith's *Wealth of Nations* in Shanghai under the title *Elements of Wealth* ('*Yuan Fu*'). In his commentary on this translation, Yan clearly stated that the principles of economics advocate free competition, are against state intervention and limit the scope of state involvement in those tasks that are not suited for the private sector. However, most Chinese intellectuals, including Yan himself, accepted the theory of liberal economics because of its contribution to the recovery of the power of the nation (Schwartz 1964).

However, the principles of liberal economics do not appear to have contributed much to the modernization of China. Late 19th-century reformers had to fight against the obsolete bureaucracy of the Qing dynasty. As was typical of revolutions in the 20th century, the social dimension of the Chinese revolution increased in significance with the passage of time. Democrats and liberals worked together on the cultural front of the 4 May Movement (1919). However, this collaboration soon broke down, since democrats shifted their position to that of Communist revolutionaries and began to attack liberals as 'bourgeois intellectuals'.

The ideology of Western socialists and social reformers was introduced by Sun Yatsen (1866–1925) through his 'Three People's

Doctrines'. Sun regarded Western capitalism as the root cause of the social problems in the West and searched for an alternative route towards economic development for China. He recognized Henry George's idea of land nationalization and the German socialist idea of capital regulation. After experiencing the state of anarchy that followed the Xinhai Revolution (1911), he sympathized with the Russian Revolution and led his Nationalist Party, the Guomingdang, in cooperation with the Communists.

Period of the Republic of China

Despite continued struggles among the warlords and an unstable security environment in both domestic as well as external affairs, the period of the Republic of China (1912–49) marked the emergence of economic academism in China. Most of the renowned universities of today originated in this period, and specialized economists, some of whom were educated in the United States, Europe and Japan, began to teach there. There were 16 Chinese publications on economics in the decade following Yan's translation of Adam Smith's *Wealth of Nations;* this number increased to 20 between the 1911 revolution and the 4 May Movement. It further increased to 228 in the decade following 1919 and to 1,116 after 1929 (Shanghaishi 2005, pp. 114–15).

The Chinese Economic Society was established in 1923, and after a decade its membership amounted to c. 600. In 1930, it launched the quarterly journal *Jinngjixue Jikan* in Shanghai. Ma Yinchu (1882–1982), a Ph.D. holder from Columbia University who had taught economics at Beijing University since 1915, was its president. He served the Guomingdan government as its economic advisor and published his views on the currency problems, banking and public finance in China. The Chinese economists of this period actively participated in policy discussions, such as the currency reforms of 1936, financial problems and industrial development plans.

However, it was the problem of agriculture that most concerned Chinese economists. A large-scale research project in rural economy headed by Chen Hansheng (1897–2004) gave birth in 1933 to the Research Forum in Chinese Rural Economy. This forum gathered a membership of about 500 members and trained economists who continued their research activity in the post-1949 period. The most prominent member among them was Xue Muqiao (1904–2005), who edited *Rural China* ('*Zhongguo Nongcun*') from 1934.

Social scientists influenced by Marxism eagerly discussed the nature of existing Chinese society (1929–1931). This debate contained a political element since those who supported the Chinese Communist Party (CCP), which was founded in 1921, regarded Chinese society to be a semi-feudal and semi-colonized society, whereas the Trotskyists emphasized the dominance of the capitalistic elements. Such debates on the nature of Chinese social history and its periodization (1931–3) and on the Asiatic mode of production continued in the field of economic history.

Marxist Monopoly Under the PRC

The People's Republic of China (PRC) started in 1949 with the programme of the 'New Democracy' that was to be based on the alliance between Communists and democrats from all sections of the society. The government requested non-resident Chinese scholars to participate in the reconstruction of China.

Ma Yinchu, who was exiled to Hong Kong as a result of a dispute with the Guomingdang government, returned to take over as the president of Beijing University. Initially, several of his colleagues were those who had been educated in American universities. Thus, in the beginning of the PRC, universities in China had non-Marxian economists on their staff. However, the socialist reconstruction of academic system based on the Soviet–Russian model, and the intensifying confrontation with the United States, soon deprived 'bourgeois economists' of freedom. Abridged translations of Russian textbooks pertaining to Marxian economics became the standard education materials. In 1957, when the CCP declared a liberal policy towards intellectuals with the

appealing phrase 'Let a hundred flowers blossom', Ma proposed his idea of population restraint to the People's Congress of the PRC. This offended Mao Zedong's positive view of population growth. The ensuing continuous attacks on 'Malthus in China' signalled the expulsion of non-Marxian ideas from the academic world under the PRC.

According to the original concept of the New Democratic Economy, the development of capitalism in China, except for 'monopoly capital', was to be welcomed as the basis for initiating future socialist transformations. However, in 1953 the success of the agrarian reforms motivated Mao to practise 'the solution to the problem of ownership'. Through the socialization of the ownership of the means of production, a Soviet Russian-type of planned economy was established in the sectors of industry and commerce during 1953–6. This was followed by the establishment of people's communes in the rural areas in 1958.

Reform Economists in China

The first criticism levelled against a centrally planned economy also emerged in the years of 'Let a hundred flowers blossom'. In 1956 and 1957, Sun Yufang (1908–83) proposed an economic model of decentralization with the use of profit targets in the management of manufacturing sector. Sun was a Marxian economist who had studied in Moscow. He grounded his proposal on the validity of the 'law of value' in a socialist economy, which is distinguished from the 'law of market'. In this respect, the views of Gu Zhen (1915–74) were more progressive, in that he openly criticized the abolition of the market mechanism under socialism. During the wave of the Anti-Rightist Struggle that occurred during the latter half of 1957, Sun and Gu were labelled 'revisionist' and 'bourgeois rightist' respectively.

Chinese economists were aware of the shortcomings of a Russian-type planned economy and the need for reform. However, the ideological rejection of the 'material interest' as a tool of 'revisionists' prevented the introduction of reforms in the management system of state-owned enterprises (SOEs). Ideological politicians stuck to the appeal to 'spiritual incentive'. Reforms were then directed towards an administrative decentralization, in which powers and benefits were divided among various administrative organs.

It was only after the declaration of the end of the Great Cultural Revolution (1966–1976) and with Deng Xiaoping (1904–1997) taking over the leadership of China that the damage caused by excessive decentralization and the need for management reforms were seriously taken into consideration. After the strategic decision of the CCP for economic reforms and an 'open door' policy, China implemented various policies such as the creation of special economic zones and township and village enterprises as well as the approval of private enterprises and households contracting in agriculture. Under the concept of the 'planned commodity economy' (1984), the market economy was theoretically subordinate to the planned economy. The existence of private sectors was legitimized by the theory of the 'early stage of socialism' (1987). At last, in the 1990s, by the definition of the 'socialist market economy' (1992), the private sector was clearly approved as the main and normal element of Chinese socialist economy.

A group of veteran economists, namely, Xue Muqiao, Du Rensheng (born 1913), Yu Guangyuan (born 1915), Liao Jili (1915–93), Lieu Guogang (born 1923), and others contributed to the transformation of the concept of 'socialist economy'. In the early 1980s, they re-examined the orthodox and heterodox texts of Marxism, studied reform economics of former socialist eastern European countries, and endeavoured to draw conclusions from the empirical research on agriculture and manufacture sectors. They formed the 'theory of the socialist commodity economy'.

After Mao's death and the end of the Great Cultural Revolution, academic economists soon regained their energy. The Chinese Academy of Social Sciences (CASS) was established in 1977. The oldest Shanghai Economics Society, whose origin can be traced back to 1950, resumed its activities in 1978. In the same year, the Chinese Research Forum of Overseas Economics was established and began to work for the diffusion

of the 'Western' (non-Marxian) economics among Chinese economists.

In the 1980s, Chinese economists recovered their communications with the world community of economists. The government invited renowned Western economists to academic conferences pertaining to the economic reforms in China. It began to send young people to the graduate courses of top Western universities, and encouraged them to assimilate advanced analysis of modern economics. By the mid-1990s, China already had a group of talented economists who could analyse economic reforms in China in a manner similar to the Western (non-Marxian) economists. In the fields of research, economic teaching, and policymaking, the activities of non-Marxian economists became more significant with each passing year. Thus, the monopoly of the Marxian economists was broken.

Present Situation of Economics in China

The ideological/political control exercised by the CCP over Chinese intellectuals had been considerably reduced at the outset of the 21st century. Economists in China can now keep themselves abreast of the latest developments in the field of economics. However, the following three features are noteworthy when compared with economics in other countries.

The first is the peculiar position of Marxian economics in China. At present, it is clear that Marxian economics is just a sub-area in the whole gamut of research activities undertaken by Chinese economists. It is therefore symbolic that the Marxian economists organized themselves into a society named the Chinese Forum for the Study of *Capital* (founded in 1981). However, Marxian economics still influences society by two privileged routes. One is that Marxian economics continues to be an obligatory course of political economy (*Zhengzhi Jingjixue*) in most Chinese universities. It is virtually a part of the political education imposed on academicians. The other route is the ideological function for the ruling CCP. The CCP needs Marxian economists to defend its policy on ideological grounds.

The second noteworthy feature of Chinese economics is the focus on institutional economics and political economy. Leading economists of the post-Great Cultural Revolution generation such as Lin Yifu (born 1952) and Fan Gang (born 1953) adopted the framework of institutional economics. Lin attributed the success of the Chinese economy after the implementation of the 'open door' policy to the switch of the development strategy and the institutional reforms accompanying it. Fan provided an analysis of the incremental reforms in China by applying the public choice approach. The theories pertaining to modern institutional economics – transaction cost theory, property rights theory, contract and corporate governance theory, and comparative institutional analysis – are widely accepted by Chinese economists.

Lastly, a new divide between the supporters of the prevailing liberal policy and its critics emerged in 2004, and a debate between these two groups has continued since then. First, Lang Xianping (born 1956), a professor at the Chinese University in Hong Kong, attacked managers of the firms whose stocks were newly listed on the stock market. They were charged with smuggling national property by the application of various techniques such as management buyouts. His attack on the privatization policy encouraged economists who were concerned about the increasing inequality in society and diminishing state intervention. They criticized over-hasty privatization and demanded a policy that would enhance the level of equality in society. They stressed the need to implement reforms in the field of social policy, and rejected the unconditional integration of the Chinese economy within the global market. Liberal economists, who stressed efficiency, rebutted them. Another group of economists declared themselves as taking a middle-of-the-road position. The government is said to have attentively followed the debate.

See Also

► Chinese Economic Reforms
► Culture and Economics

Bibliography

Fawcett, H. 1883. *Fuguoce* [Policies for the wealth of nations]. Translation by Wang Fengzao under the supervision of W.A.P. Marten of *A manual of political economy*, London: Macmillan, 1863. Shanghai: Shanghai Meihua Shuguan.

Hu Jichuang. 1988. *A concise history of Chinese economic thought*. Beijing: Foreign Language Press.

Jevons, W.S. 1886. *Fuguo Yangmince* [Policies for the wealth of nations and support of people]. Translation by J. Edkins of *Primer of political economy*, London: Macmillan, 1878. Zongshuiwusishu.

Schwartz, B.I. 1964. *In search of wealth and power: Yen Fu and the West*. Cambridge, MA: Harvard University Press.

Shanghaishi Shehuikexue Lianhehui (ed.). 2005. *20 Shiji Zhongguo Shehuikexue Lilun Jingjixue* [Social sciences in the 20th century China, economic theory]. Shanghai: Shanghai Renmin Chubenshe.

Trescott, P.B. 2006. *Jingji Xue: History of introduction of western economic ideas into China 1850–1950*. Hong Kong: Chinese University Press.

Wei Yuan. 1843. *Haiguo Tuzhi* [Geography of the maritime countries]. Yangzhou.

Wu Jinglian. 2005. *Understanding and interpreting Chinese economic reform*. Mason: Thomson Higher Education.

Yu Yingshi. 1987. *Zhongguo Jinshi Zongjiao Lunli yu Shangren Jingshen* [Religious ethics and spirit of merchants in early modern China]. Taipei: Lianjian Chuben.

Zhao Jing (ed.). 1991–8. *Zhongguo Jingji Sixiang Tongshi* [Complete history of Chinese economic thought], vols. 4. Beijing: Beijing Daxue Chubanshe.

Zhao Jing (ed.). 2004. *Zhongguo Jingji Sixiang Tongshi Xuji: Zhongguo Jingji Jindai Sixiangshi* [Complete history of Chinese economic thought continued: History of modern Chinese economic thought]. Beijing: Beijing Daxue Chubanshe.

Zhongguo Dabaike Quanshu Zongbianji Weiyuanhei Jingjixue Bianji Weiyuanhei (ed.). 1998. *Zhongguo Dabaike Quanshu, Jingjixue 1* [Great encyclopaedia of China, economics 1]. Beijing: Zhongguo Dabaike Quanshu Chubenshe.

China's One Child Policy

Lisa Cameron and Xin Meng

Keywords

Aging; China; Family planning; Gender; One child policy

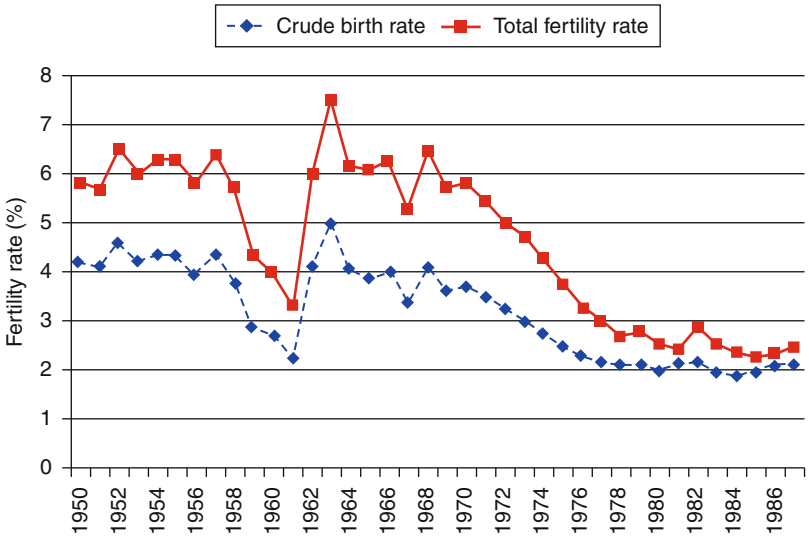
JEL Classification

J10; J16; O12; O29; O53; P21

Background

When the Communist Party came to power in China the country was recovering from the Second World War and its civil war. The war-torn country desperately needed people to rebuild the nation. Consequently, the country's population policy at that time was to encourage a high birth rate. Along the lines of the Soviet model, China awarded mothers of many children "Mother-Hero" status and, like most post-Second World War western countries, China had a baby boom, with the total fertility rate hovering around six births per mother. This period of high fertility lasted until the great famine in the late 1950s. The three-year famine significantly reduced the birth rate, but soon after population growth bounced back to over and above the pre-famine level (sees Fig. 1). China's total fertility exceeded six births per mother throughout the 1960s (Banister 1987).

Low agricultural productivity, economic stagnation and economic isolation as a result of the Cold War, coupled with vivid memories of the Great Famine, led to concern that China might again not be able to feed its growing population. In the early 1970s, in the middle of the Cultural Revolution, a philosophical 'debate' over the possibility of a Malthusian Population Trap was initiated, and soon after, at the end of 1973, the policy of 'Later, Longer, and Fewer' was introduced (Center for Population Studies, CASS 1986; Peng 1991; Feeney and Wang 1993; Cai 2010; Ebenstein 2010). The policy encouraged couples to get married later, have longer periods in between children, and have fewer children. The policy had a significant impact on the birth rate, especially amongst the urban population. During the course of the 1970s the total fertility rate fell from above 5 to just around 2 (Wang 2011). In January 1978 a new policy of 'One is the Best and Two is the Most' and 'Reward Having One Child and Punish Having Three' was introduced. This was soon followed by the introduction of the 'One



China's One Child Policy, Fig. 1 China's fertility rates, 1950–1987 (Sources: The data for years 1950 to 1972 are from Banister (1987) and for the years from 1973 to 1978 are from Feeney et al. (1989))

C

Child per Couple' policy (hereinafter the One Child Policy (OCP)) at the second meeting of the fifth People's Congress in June 1979 (see, for example, Center for Population Studies, CASS 1986; Peng 1991).

Although the OCP was originally intended to cover the country as a whole, by and large, rural areas have always allowed a second birth if the first child is a girl (Peng 1991). The proportion of the population with rural household registration (*hukou*) was more than 80% in 1980 and was still above 70% in 2010. Some rural areas over some periods have even allowed three children. In addition, there are different rules for minority groups, which are subject to much looser restrictions. In urban areas, however, the policy has been strictly enforced since it was introduced (Kane and Choi 1999; Zhang and Sturm 1994). Those who obey the policy are rewarded financially while those who violate the policy are subject to fines and their children face higher fees for accessing education and health services. In some cases children are denied these services (Peng 1991; Zhang and Sturm 1994).

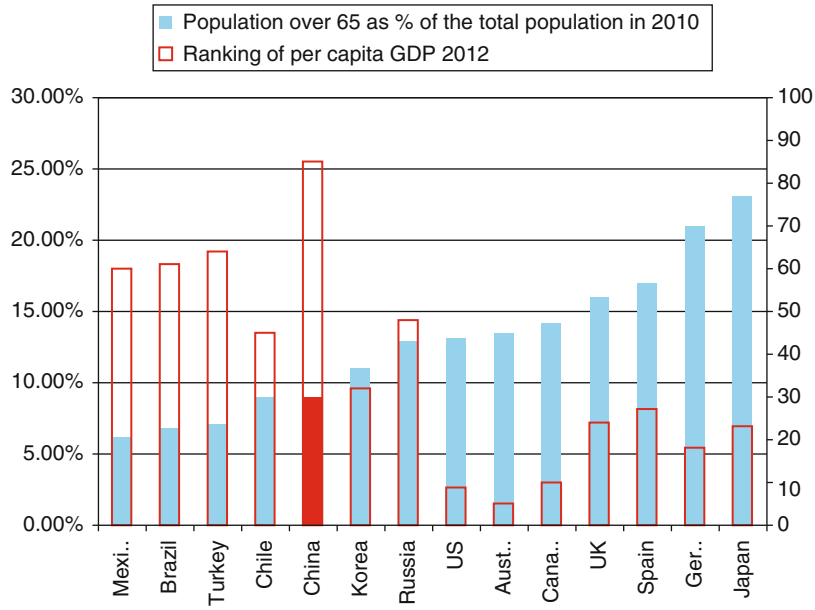
Impacts

Population Size

The ultimate objective of the introduction of the OCP was to control population size. To be specific, the goal was to keep the total population at no more than 1.2 billion by 2000 (Wang et al. 2012). During the past few years, there has been heated debate in the literature over the magnitude of the impact of the policy on population size in China.

The official Chinese government claim is that some 400 million births were prevented due to the OCP. According to Wang et al. (2012), the original '400 million' figure came from a simple extrapolation conducted by Yang et al. (2000) to project what the crude birth rate would have been if it had continued to decline as it did from 1950 to 1970. Wang et al. (2012) challenge the 400 million estimate for two reasons: (1) they argue that Yang et al.'s (2000) original projection used the wrong counterfactual. (They did not take into account the fact that between 1970 and 1979, before the introduction of the OCP, China's fertility rate had already dropped by 50%.); and (2) many countries which had similar fertility rates to China in 1970 also experienced significant drops in fertility in the absence of the OCP. For example, Thailand and China have had almost identical fertility trajectories since the 1980s, but Thailand did not adopt a one child policy.

Using the actual Chinese fertility rate in 1979 (just before the introduction of the OCP) and the fertility trends in 16 other countries, which had similar fertility rates to China in the year 1970,



China's One Child Policy, Fig. 2 Share of population over 65 and economic development (Source: Aging population data are from OECD StatExtracts: http://stats.oecd.org/Index.aspx?DataSetCode=ALFS_SUMTAB, while per capita GDP data are from World Economic Outlook Database – October 2013, International Monetary Fund)

Wang et al. (2012) use a Bayesian model to project China's counterfactual fertility trend. Their projection is that fertility would have continued to decline after 1980 and would have fallen to 1.5 children per woman, as currently observed. In other words, without the OCP China would have had the current population size anyway.

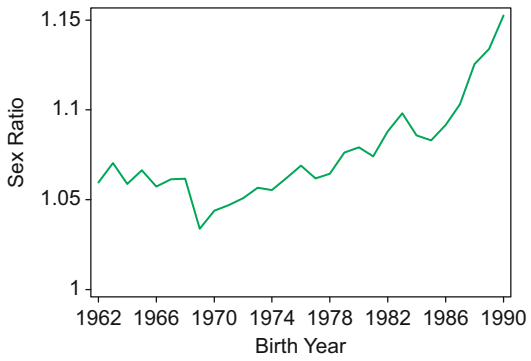
The main issue in this debate is whether the 16 other countries are a credible counterfactual. Most of the 16 countries have very different income levels, institutional structures, and cultures to China and all these potentially contribute to fertility behaviour. In addition, Wang et al. (2012) acknowledge that the proportion of households that have only one child in China is very high and admit that the proportion would never have reached the high level observed without the OCP. If this is the case, then the OCP must have played some role in curtailing population growth.

Population Aging

If one accepts that the OCP did reduce the fertility rate, then population aging is one of the significant consequences of the reduction in births over more than three decades. Concomitant increases in life expectancy have further exacerbated population aging. In 1980 around 5.8% of China's population were aged 65 and over. Three decades later, this ratio had increased to 9%. Although, relative to many developed countries, this is not an especially high proportion (for example, in the same year the USA and Japan had, respectively, 13% and 23% of their populations over 65), China is aging with much lower income levels.

Figure 2 compares China's aging population ratio with several other countries. China is ranked below the median level, but all the countries listed have much higher per capita income levels than China. For example, in 2012 Chinese per capita GDP was ranked 85th (US $6,569) among the 183 countries, while Chile, a country with the same aging population ratio as China, was ranked 45th in per capita GDP (US$ 16,834). The country closest to China's income level is Turkey, with per capita GDP ranked 64th (US$ 10,745), and whose aging population ratio was much lower than China, at 7.1%.

Lower income with a rapidly aging population inevitably creates a financial burden on the government. This is because aged care, especially medical care, is costly. China's Health and Retirement Survey team report that a large fraction of China's elderly have physical health limitations (CHARLS Research Team 2013). The problem is even more serious when one considers the



China's One Child Policy, Fig. 3 Sex ratios by birth cohort (Source: Replicated from Fig. 1 in Li et al. (2011). Calculated based on the 1990 Chinese population census (1% sample))

proportion of households in which the responsibility for looking after elderly parents and/or grandparents falls on only one child (an average of 40% of households across the country as a whole and 86% of households with urban residency (*hukou*); Wang et al. (2012)). China has traditionally relied heavily on families to provide aged care. Even today the Elderly Rights and Security Law states that adult children have financial and emotional responsibility for their elderly parents (Zhang 2000). Previously the parental old age care responsibilities could be shared among siblings. However, this will not be possible for the single-child generation. Thus additional social care and social expenditure will be called upon.

Gender Equality

Figure 3 shows China's sex ratio from 1950 (almost three decades prior to the introduction of the OCP) until 1990. For most of 1950 to the late 1970s China's sex ratio remained very close to 106. Since the introduction of the OCP in 1979 the sex ratio has increased to over 113. That is an extra 13 boys for every 100 girls born.

The OCP, in concert with the strong cultural preference for sons and the availability of gender selection technology such as ultrasound scans and induced abortion, is widely believed to be the major reason for the increase in the sex ratio (Ebenstein 2010; Das Gupta 2005; Zeng et al. 1993), and various studies have sought to quantify the policy's impact. It has been estimated that in excess of 40 million women are 'missing' in China (Klasen and Wink 2002). Li et al. (2011) use minority groups who are not subject to the OCP as a control group and estimate that the policy led to 4.4 extra boys per 100 girls in the 1980s, and to 7.0 extra boys per 100 girls for the period 1991–2005. This corresponds to approximately 94% of the total increase in the 1980s and about 55% for 1991–2005. Bulte et al. (2011) use a similar methodological strategy to identify the policy's impact and attribute 50% of the missing women to the policy. As with the policy's impact on population size, and because China was experiencing substantial change over this period with economic and other reforms accompanying the introduction of the OCP, there is some ongoing debate as to the extent to which the OCP is responsible. Other explanations put forward to explain increasing sex ratios include increasing gender wage gaps due to changes in centrally determined agricultural crop procurement prices (Qian 2008) and land reform (Almond et al. 2013). Oster (2005) argued that Hepatitis B infection is associated with an increased probability of giving birth to a son and so could potentially explain the sex ratio imbalance. This point was, however, disputed by Das Gupta (2008) on the basis of Lin and Luoh (2008)'s finding that Hepatitis B only increased the probability of having a boy fractionally in a large medical data set for Taiwan. Oster et al. (2010) subsequently also found no increases in the probability of having a boy associated with Hepatitis B prevalence in China and conceded that the disease did not drive the increases in China's sex ratio.

The sex ratio imbalance has serious consequences. Guilmoto (2012) estimates that the number of prospective grooms will exceed the number of prospective brides by more than 50% for at least three decades. The number of unmarried men at age 50 is estimated to peak at 15% in 2055. The excess of men has resulted in women marrying older men and a lesser education gap between men and women. To increase marriageability, families with boys are increasing their savings (Wei and Zhang 2011). Some men will not be able to find partners though, regardless of how much

savings they have. Das Gupta et al. (2010) project that unmarried males will likely be concentrated in poorer provinces with lower ability to provide social protection to their citizens and predict that such geographic concentration of unmarried males could be socially disruptive. Indeed, Edlund et al. (2013) find that increases in the sex ratio account for almost 15% of the recent large increases in crime rates in China.

As egregious as sex-selective abortion and female infanticide are, females who survive beyond infanthood are benefitting from their scarcity in some ways. The high sex ratio gives women greater bargaining power within the household, which consequently results in them bearing fewer children and investing more heavily in children's health (Porter 2007). Higher sex ratios are also associated with married women being less likely to live with their in-laws Li (2012). Lee (2011) finds that the OCP has reduced the gender education gap, as girls who are only children are no less likely to be educated than boys who are only children. Ultimately son preference will diminish as parents prefer to have a girl who they will be able to marry off rather than a boy with worse marriage prospects, Edlund (1999).

Behavioural Consequences

Over the past 30 years the OCP is felt to have had a marked impact on the behaviour of the one-child generation. There are claims in the media as well as some psychologists that this new generation could be different (see, for example, Lee (1992), Fan (1994) and Wang et al. (1998)). People are concerned that parents have not been teaching their only children traditional values and that the children of the one-child generation are more self-centred and less cooperative than previous generations. This sentiment is captured in the common phrase that the only child is the 'little emperor' of the family. As the first cohorts of the one-child generation entered the labour market, employers started to include phrases such as 'no single children need apply' in job ads (see Chang 2008).

Empirical evidence of these behavioural differences is limited. Results of studies that compare only children with others produce mixed results (Chen and Goldsmith 1991; Falbo and Poston 1993; Shen and Yuan 1999; Wang et al. 2000; Liu et al. 2005). These comparisons are, however, unlikely to be a comparison of like with like, as they ignore that families who are able to have more than one child under the OCP are often substantially different from families who have had to comply with the policy. A comparison of people (whether only children or otherwise) born just before and just after the introduction of the OCP in 1979 avoids this problem, as for the majority of the population having more than one child was no longer a choice after the policy's introduction. Such a comparison was conducted by the authors (and their co-authors) in Beijing in 2010. Personality surveys and economic experiments designed to elicit the extent of pro-social and other behaviours revealed strong behavioural differences, largely in line with the anecdotal reports – the OCP was found to have produced a less trusting, less trustworthy, more risk-averse, less competitive, less conscientious, more pessimistic and more neurotic generation (Cameron et al. 2013).

Economic Consequences

The overall economic impact of the OCP is very difficult to determine. On the one hand, the policy has reduced the country's stock of labour and generated a heavy burden on the younger generation, as only children struggle to support their parents and four grandparents financially. Entrepreneurial spirit has also possibly been sapped as the result of the generation's reduced willingness to take risks (Cameron et al. 2013). On the other hand, the policy seems to have stimulated saving, although recent studies have found that the general equilibrium effect of the OCP on savings could be small (Banerjee et al. 2011, 2014; Choukhmane et al. 2013). Investment in education for both genders has also increased (Rosenzweig and Zhang 2009). Zhu et al. (2013) argue that the demographic changes caused by the OCP may not harm China's longterm growth. Their model, which focuses largely on fertility decreases coupled with increased educational attainment, estimates that by 2025 China's GDP will be about 4% higher than it would have been without the OCP.

Potential Implications of Policy Changes

In late 2013 China announced that it was relaxing its OCP. Couples in which at least one is an only child will now be allowed to have two children.

The government predicts that the relaxation of the policy will be associated with a significant increase in the total birth rate in the initial stage to above 1.8, but in the long run the birth rate is estimated to settle at around 1.6 to 1.7 (*People's Daily* 2013). The relaxation of the OCP is unlikely to have a large impact on population size for the following reasons:

1. Currently, 70% of the Chinese population has rural household registration (*hukou*), and for these people the OCP was never binding. The majority of rural *hukou* households have two or more children and hence for them the new policy will not have an effect.
2. For the remaining urban *hukou* population (30%) the past thirty years of economic growth has likely changed people's fertility preferences. This change in fertility preferences reflects that the majority of urban women are going to university and hence delay their first birth and have a shorter period in which they wish to bear children. Further, the cost of rearing children has risen sharply in cities.

Many demographic studies support the second point. For example, Cai (2010) uses differences in the implementation of the family planning policy in two economically fairly developed provinces, Zhejiang and Jiangsu, as a natural experiment to examine the effect of the OCP versus a policy which allows for two children on the total fertility rate. He finds that despite the differences in the family planning policy, the total fertility rate is similar across the two provinces, suggesting that economic growth has dominated individuals' fertility behaviour.

If the relaxation of the OCP does not have a large effect on the population size in the long run, it will also not have a large effect on the aging of the population. The official government view however is that it will improve the age structure of the Chinese population. The *People's Daily* reports that the change in policy will gradually reduce the share of the aging population, so that by 2100 34.3% of the population are elderly, compared to 39.6% of the population in the absence of any policy change (*People's Daily* 2013).

The changes in the OCP promise to reduce the sex ratio and some of the associated consequences described above. Parents living in urban areas who were subject to the strict one-child version of the policy and who may have aborted a first child daughter are now likely to resort to sex selectivity only if the first two children are daughters. As discussed above, most rural areas already allowed couples whose first child was a daughter to have a second child. This is often referred to as the '1.5 rule'. Zeng (2007) shows that sex ratios are higher in 1.5 child areas than 2 child areas. He found the sex ratio at birth in 2000 in areas where there are '1.5 children' rules to be 119.7, compared to 108.3 in 2 child policy areas (see also Goodkind 2011).

The decline in the sex ratio at birth will, with time (approximately two decades), relieve the stress in the marriage market and reverse many of the phenomena described above – men will start being able to marry at younger ages. Female labour supply may increase as a result of reduced female bargaining power. There may be adverse effects on resources devoted to children. Some changes are, however, unlikely to be reversible, as cultural norms will have shifted irretrievably. This is probably true of son preference, which has weaker foundations in a modern economy in which wages are rising in the female-dominated service sector relative to manufacturing and blue collar jobs. Market forces are also likely to counter any reductions in investment in the human capital of girls. A smaller pool of unmarried men is likely to have positive impacts on social cohesion.

The extent to which the behavioural impacts of the policy can be ameliorated by the policy's relaxation will depend on its impact on fertility. If one-child families remain the norm, the behavioural tendencies identified by Cameron et al. (2013) can be expected to persist, unless they are offset by specific educational and institutional efforts. The increasing marketization of society is further eroding pro-social values, so

those who wish for a return to a society of traditional values are likely to be disappointed.

See Also

- China, Economics in
- Chinese Economic Reforms
- Demographic Transition
- Maoist Economics
- Population Ageing

Bibliography

Almond, D., H. Li, and S. Zhang. 2013. *Land reform and sex selection in China*. NBER Working Paper No. 19153, 1–52.

Banerjee, A., X. Meng, and N. Qian. 2011. *The life cycle model and house-hold savings: Micro evidence from urban China*. Unpublished manuscript, Yale University.

Banerjee, A., X. Meng, T. Porzio, and N. Qian. 2014. *Aggregate fertility and household savings: A general equilibrium analysis using micro data*. NBER Working Paper No. 20050.

Banister, J. 1987. *China's changing population*. Palo Alto: Stanford University Press.

Bulte, E., N. Heerink, and X. Zhang. 2011. China's one-child policy and 'the mystery of missing women': Ethnic minorities and male-biased sex ratios. *Oxford Bulletin of Economics and Statistics* 73(1): 21–39.

Cai, Y. 2010. China's below-replacement fertility: Government policy or socioeconomic development? *Population and Development Review* 36(3): 419–440.

Cameron, L., N. Erkal, L. Gangadharan, and X. Meng. 2013. Little emperors: Behavioral impacts of China's one child policy. *Science* 339: 953–957.

Center for Population Studies, Chinese Academy of Social Sciences & Editorial Board of the China's Population Yearbook. 1986. The major events of China's population activities. In *China's Population Yearbook (1985)*, 1263–1288. Beijing: Chinese Social Sciences Press.

Chang, L. 2008. *Factory girls: From village to city in a changing China*. New York: Spiegel and Grau.

CHARLS Research Team. 2013. *Challenges of population aging in China: Evidence from the National Baseline Survey of the China Health and Retirement Longitudinal Study (CHARLS)*, unpublished report: http://charls.ccer.edu.cn/uploads/document/public_documents/application/Challenges-of-Population-Aging-in-China-final0916.pdf.

Chen, J., and L.T. Goldsmith. 1991. Social and behavioral characteristics of Chinese only children: A review of research. *Journal of Research in Childhood Education* 5(2): 127–139.

Choukhmane, T., N. Coeurdacier, and K. Jin. 2013. *The one-child policy and household savings*. LSE Working Papers, London School of Economics.

Das Gupta, M. 2005. Explaining Asia's 'missing women': A new look at the data. *Population and Development Review* 31: 529–535.

Das Gupta, M. 2008. Can biological factors like Hepatitis B explain the bulk of gender imbalance in China? A review of the evidence. *World Bank Research Observer* 23(2): 201–217.

Das Gupta, M., A. Ebenstein, and E. Sharygin. 2010. *China's marriage market and upcoming challenges for elderly men*. World Bank Policy Research Working Paper No. 5351. Washington DC: World Bank.

Ebenstein, A. 2010. The 'missing girls' of China and the unintended consequences of the one child policy. *Journal of Human Resources* 45(1): 87–115.

Edlund, L. 1999. Son preference, sex ratios, and marriage patterns. *Journal of Political Economy* 107(6): 1275–1304.

Edlund, L., H. Li, J. Yi, and Z. Junsen. 2013. Sex ratios and crime: Evidence from China. *Review of Economics and Statistics* 95(5): 1520–1534.

Falbo, T., and D. Poston. 1993. The academic, personality, and physical outcomes of only children in China. *Child Development* 64(1): 18–35.

Fan, C. 1994. A comparative study of personality characteristics between only and non-only children in primary schools in Xian. *Psychological Science* 17(2): 70–74 (in Chinese).

Feeney, G., and F. Wang. 1993. Parity progression and birth intervals in China: The influence of policy in hastening fertility decline. *Population and Development Review* 19(1): 61–101.

Feeney, G., F. Wang, M. Zhou, and B. Xiao. 1989. Recent fertility dynamics in China: Results from the 1987 one percent population survey. *Population and Development Review* 15(2): 297–322.

Goodkind, D. 2011. Child underreporting, fertility, and sex ratio imbalance in China. *Demography* 48: 291–316.

Guilmoto, C. 2012. Skewed sex ratios at birth and future marriage squeeze in China and India, 2005–2100. *Demography* 49: 77–100.

Kane, P., and C. Choi. 1999. China's one child family policy. *British Medical Journal* 319(7215): 992–994.

Klasen, S., and C. Wink. 2002. A turning point in mortality? An update on the number of missing women. *Population and Development Review* 28: 285–312.

Lee, L. 1992. In *Child care in context: Cross cultural perspectives*, ed. M. E. Lamb, K. Sternberg, 355–392. Hillsdale: Lawrence Erlbaum.

Lee, M.-H. 2011. The one-child policy and gender equality in education in China: Evidence from household data. *Journal of Family Economics Issues* 33: 41–52.

Li, Q. 2012. The effects of sex ratio imbalance in China. In *Population Association of America Meeting*, 3–5 May, San Francisco.

Li, H., J. Yi, and J. Zhang. 2011. Estimating the effect of the one-child policy on the sex ratio imbalance in

China: Identification based on the difference in differences. *Demography* 48: 1535–1557.
Lin, M.-J., and M.-C. Luoh. 2008. Can hepatitis B mothers account for the number of missing women? Evidence from three million newborns in Taiwan. *American Economic Review* 98(5): 2259–2273.
Liu, C., T. Muakata, and F. Onuoha. 2005. Mental health condition of the only-child: A study of urban and rural high school students in China. *Adolescence* 40: 831–845.
Oster, E. 2005. Hepatitis B and the case of missing women. *Journal of Political Economy* 113(6): 1163–1216.
Oster, E., G. Chen, X. Yu, and W. Lin. 2010. Hepatitis B does not explain malebiased sex rations in China. *Economics Letters* 107: 142–144.
Peng, X. 1991. *Demographic transition in China*. Oxford: Clarendon Press.
People's Daily. 2013. *The adjustment in family planning policy will change share of ageing population, labour and public resources demand and supply (in Chinese)*. http://paper.people.com.cn/rmrb/html/2013-11/18/nw.D110000renmrb_20131118_1-02.htm.
Porter, M. 2007. *The effects of sex ratio imbalance in China on marriage and household bargaining*. Working Paper. Chicago: University of Chicago.
Qian, N. 2008. Missing women and the price of tea in China: The effect of sexspecific earnings on sex imbalance. *Quarterly Journal of Economics* 123: 1251–1285.
Rosenzweig, M., and J. Zhang. 2009. Do population control policies induce more human capital investment? Twins, birth weight and China's 'one-child' policy. *Review of Economic Studies* 76(3): 1149–1174.
Shen, J., and B.J. Yuan. 1999. Moral values of only and sibling children in mainland China. *Journal of Psychology: Interdisciplinary and Applied* 1331: 115–124.
Wang, F. 2011. The future of a demographic overachiever: Long-term implications of the demographic transition in China. *Population and Development Review*, 37(suppl.): 173–190.
Wang, Q., M.D. Leichtman, and S.H. White. 1998. Childhood memory and selfdescription in young Chinese adults: The impact of growing up an only child. *Cognition* 69: 73–103.
Wang, D., N. Kato, Y. Inaba, T. Tango, Y. Yoshida, Y. Kusaka, Y. Deguchi, F. Tomita, and Q. Zhang. 2000. Physical and personality traits of preschool traits in Fuzhou, China: Only children vs sibling. *Child: Care, Health and Development* 26(1): 49–60.
Wang, F., Y. Cai, and B. Gu. 2012. Population, policy, and politics: How will history judge China's one-child policy? *Population and Development Review*, 3(suppl.): 115–129.
Wei, S., and X. Zhang. 2011. Sex ratio imbalances stimulate savings rates: Evidence from the missing women in China. *Journal of Political Economy* 119(3): 511–564.
Yang, K., S. Chen, and J. Wei (eds.). 2000. *China's birth planning: Benefits and inputs (in Chinese)*. Beijing: People's Press.
Zeng, Y. 2007. Options for fertility policy transition in China. *Population and Development Review* 33: 215–246.
Zeng, Y., T. Ping, B. Gu, Y. Xu, B. Li, and Y. Li. 1993. Causes and implications of the recent increase in the reported sex ratio at birth in China. *Population and Development Review* 19: 283–302.
Zhang, K.D. 2000. *Textbook on the elderly law*. Beijing: China Encyclopedia Press.
Zhang, J., and R. Sturm. 1994. When do couples sign the one-child certificate in urban China? *Population Research and Policy Review* 13(1): 69–81.
Zhu, X., J. Whalley, and X. Zhao. 2013. *Intergenerational transfer, human capital and long-term growth in China under the one child policy*. NBER Working Paper No. 19160.

Chinese Economic Reforms

Loren Brandt and Thomas G. Rawski

Abstract

Why did China's modest reforms unleash an enormous boom? Three decades of socialist planning created vast untapped potential. China captured this potential by focusing on 'big reforms' linked to incentives, markets, prices, mobility, openness and competition. Advances in these areas created sufficient momentum to overcome the drag associated with remaining distortions and institutional shortcomings. China's political economy, which incorporates substantial local autonomy, facilitated experimentation that repeatedly identified feasible reform paths. Because China's political economy delivers undesirable outcomes along with rapid growth, and because China's success is linked to unique historical circumstances, the beneficial outcomes associated with Chinese policies and institutions may be limited in time and space.

Keywords

Agricultural productivity; China, economics in; Chinese economic reform; Collectivization; Decentralization; Dual price system; Planning;

Rural–urban migration; Socialist market economy (China); Cultural Revolution (China); Great Leap Forward (China)

JEL Classifications
P3

Since the late 1970s, China's economic performance has astonished the world. Official figures show that, after adjusting for inflation, China's GDP grew at an annual rate of 9.7 per cent between 1978 and 2006, and at a rate of 8.4 per cent in per capita terms (Yearbook 2006, p. 60; National Bureau of Statistics 2006). By 2006, the Chinese economy, measured in terms of purchasing power parity, was the world's second largest, behind only the United States: per capita incomes, measured on the same basis, rose from 324 dollars to 5,772 dollars between 1978 and 2004 (Heston et al. 2006). China's new dynamism includes a major shift towards intensive growth, with productivity change, which had contributed negatively to Chinese growth between 1957 and 1978, accounting for 40 per cent of overall growth after 1978 (Perkins and Rawski forthcoming).

Reform began in the late 1970s. The impetus for modifying the plan system came from two sources: general awareness that China's neighbours were running far ahead in the economic sphere, and stagnation of living standards, especially China's persistent problems with food supply. The initial objective was to improve economic results under the system of central planning.

Initial Reform Efforts

Not surprisingly, early reform efforts focused on agriculture. Starting in 1978, household cultivation swiftly replaced collective tillage as the norm in China's vast farm sector, as hundreds of millions voted with their feet to abandon collective farming, the central feature of the people's communes.

Introduction of the household responsibility system meant that farmers could claim the fruits of extra effort for themselves. This brought an immediate multiplication of work effort, which was further encouraged by modest relaxation of restrictions on marketing and price flexibility, and by a considerable increase in procurement prices (Sicular 1995). The result was a sudden upsurge of farm production and productivity (Lin 1992). With the expansion of food supply, millions of farmers no longer needed to work the land and so began to move into non-farm employment. Improved diets raised the energy levels and hence the productivity of formerly undernourished villagers. Relaxation of efforts to enforce local selfsufficiency in favour of historic patterns of crop specialization, along with new opportunities to diversify into animal husbandry, horticulture, and aquaculture, also contributed to steep gains in farm output (Lardy 1983).

The response to agricultural reform quickly spread beyond the farm sector. Rural factories, which had enjoyed a brief boom during the Great Leap Forward of 1958–60 (a massive and chaotic push to organize villagers into communes and to transfer rural labour into steel and other industries), suffered considerable retrenchment during the 1960s, and then expanded rapidly during the 1970s. Following the revival of agriculture, collectively owned rural industry, now fortified by greater access to the cities, rising rural incomes, increased supplies of agricultural inputs, and throngs of job-seekers, bounded ahead. In addition, new freedom encouraged a wide range of non-farm self-employment and family businesses. The resulting shift out of farming initiated what eventually became a massive exodus of labour from the countryside.

The explosive response to rural reform spurred officials to press forward with urban initiatives focused on 'enlivening' state-owned enterprises. While these early measures achieved only limited progress towards their main objective, they benefited rural and urban collective industry by opening new markets as well as new sources of materials, subcontracting opportunities, and technical expertise.

As the influence of markets, price flexibility, and mobility expanded, a separate strand of reform began to move China's isolated system towards greater participation in international trade and investment. China's leaders agreed to

establish four tiny 'special economic zones' in the southern provinces of Guangdong and Fujian. Initial operations in these zones seemed directionless and inconsequential, but the arrival of ethnic Chinese entrepreneurs, most from Hong Kong and Taiwan, turned the zones into drivers of regional and eventually national growth. This novel combination of low-cost Chinese labour with the market knowledge and entrepreneurial capabilities of overseas Chinese businessmen gradually developed into an export bonanza that nudged China towards its subsequent embrace of economic globalization.

Although the limited extent of domestic reform restricted the initial response to growing openness, the buoyant prosperity of the new zones prompted cities along the coast, and eventually across the nation to clamour for access to the same tax, legal, and regulatory concessions that had powered their growth.

China's initial reforms focused on limited changes directed at specific sectors. These changes proved sufficient to accelerate growth despite the continued importance of state ownership, price controls, material-balance planning, and other key features of the socialist system. Early reform was particularly successful in removing long-standing constraints formerly imposed by limited availability of food and of foreign exchange.

Further Reforms: Expanding the Cage

During this period, China's gathering boom encouraged a growing array of jurisdictions, constituencies, and interest groups to pursue the advantages enjoyed by reform participants, including expanded managerial autonomy and access to the special economic zones. The image of China's economy as a caged bird advanced by Chen Yun, an economic specialist within the leadership group, illustrates the underlying economic thinking (Lardy and Lieberthal 1983). Chen argued that expanding the cage (reform) allows the bird to beneficially spread its wings; an overlarge cage threatens loss of control – thus the slogan 'planned economy as the mainstream, market allocation as a supplement'.

Implementation of the dual price system, which partitioned allocation of most commodities into plan and market components and allowed the distribution of afterplan residuals at increasingly flexible prices, stands as the central policy achievement of this period. The expansion of market transactions began to whittle away at longstanding barriers to mobility, which had restricted the transfer of labour, capital, commodities and ideas across administrative boundaries, with negative consequences for growth of output and productivity.

Developments in the international sphere, including the continued growth of foreign trade, the northward spread of special zones, and the expansion of foreign direct investment, now involving multinational corporations as well as overseas Chinese entrepreneurs, extended the impact of market forces. The growth of crossborder transactions and the increased presence of foreign business operations on Chinese soil intensified pressures for contract arbitration, codification of urban landuse rights and other legal and institutional reforms needed to facilitate new activities.

The main impact of these reforms fell on flows – of labour, commodities, profits, and new investments. New entrants to the workforce, for example, including college graduates, were increasingly left to find their own positions, rather than receiving job assignments from local labour bureaus. Existing stocks, including assets or employees of extant firms, especially in the state sector, were not yet exposed to the full impact of market forces. Mergers appeared, but only on a microscopic scale. Despite the enactment of bankruptcy legislation, floundering companies rarely disappeared. Nor did redundant workers face the sack, although the 'optimal labour programme', which invited managers to identify essential and surplus workers, foreshadowed the mass layoffs of the late 1990s.

Economic Reforms Since 1992: Towards a 'Socialist Market Economy'

The brief recession, triggered by efforts to quell inflation during the late 1980s, together with the

anti-reform backlash and pullback of foreign investment that followed the June 1989 suppression of popular unrest, slowed both growth and reform. The setback, however, was short. Deng Xiaoping's call for expanded reform during his southern tour of 1992, together with the Communist Party's 1992 decision to pursue a socialist market economy with Chinese characteristics gave fresh impetus and as well as new direction to economic reform.

The Party's 1992 decision replaced vague ideas of 'doing better' with a clear reform objective: a market economy in which the eventual role of the state will resemble the current circumstances of major economies such as those of France or Japan: macroeconomic management; regulation of health, environment, and so on; and strategic planning, with other functions explicitly assigned to the sphere of market determination.

Although the 1992 decision is a statement of principle rather than a description of reality, the ensuing 15 years witnessed decisive strides towards market outcomes, which we summarize in terms of four major shifts:

1. *From plan to market*: price liberalization extended beyond the substantial achievements of the first reform decade: despite significant exceptions (energy, credit, foreign exchange) supply and demand now determine most prices (Li 2006, pp. 104–7). The growing influence of market forces brought a considerable (but incomplete) hardening of budget constraints, even in the state sector. Market pressures compelled the dismissal of more than 50 million workers, most from state-owned factories. Mergers and acquisitions extended the reach of market pressures to much of China's capital stock. Barriers to the free flow of labour and goods continue to recede, and migrant workers have begun to attain normal citizenship rights in China's cities and towns. Growing expansion of wage differentials and of income inequality reflect the new prominence of market outcomes.
2. *From village to town and city* and from agriculture to industry and services. The primary sector's GDP share dropped from 27.9 per cent in 1978 to 11.8 per cent in 2006. Following the departure of 150–200 million villagers from the land, survey data indicate that the primary sector's labour force share has declined from 69.2 per cent in 1978 to 31.8 per cent in 2004 (National Bureau of Statistics 2006; Yearbook 2006, p. 58; Brandt et al. 2008).
3. *From public to private ownership*. At the start of reform, the public sector (including collectives) held nearly all China's fixed capital. The growth of private business, while rapid in percentage terms, started from a tiny base. It was only from the late 1990s that the non-public sector, swollen by the privatization of rural collective enterprises, the transfer of (mostly small and medium) state-owned firms into private hands, and the rapid expansion of direct foreign investment, began to take on a prominent role in the national economy. The share of state-owned firms in industrial output fell from 81 per cent to 55 per cent between 1980 and 1990, and to 15–35 per cent in 2005/6 (depending on the treatment of state shareholdings; see National Bureau of Statistics 2006; Perkins and Rawski 2008). The pace of change has accelerated: by 2003, the private sector's GDP share had risen to 59.2 per cent (OECD 2005, p. 125). The state sector's share in industrial output and non-farm employment during 2004/5 declined to 15.2 and 13.1 per cent (Yearbook 2006, p. 505; Brandt et al. 2008). Following lengthy reform efforts China's major banks and financial firms have begun to sell partial ownership stakes to overseas financial companies.
4. *From isolation to global engagement*. Beginning from near-autarchy during the 1960s and 1970s, China has gradually emerged as a leading participant in global trade. China's 2001 entry into the World Trade Organization (WTO) capped a gradual process of opening that has raised the ratio of combined imports and exports to GDP from under ten per cent prior to the reform to over 63 per cent in 2005 – surpassing comparable figures for all other large and populous nations (Lardy 2002; Brandt et al. 2007). China has become the world's largest recipient of foreign direct

investment, which initially clustered in manufacturing, but has recently extended into finance, property, retailing, logistics, infrastructure and R&D. Foreign firms have taken the lead in integrating China into multinational supply chains for manufacturing, research and design. Chinese firms have also begun to increase their own overseas investment in pursuit of raw materials, market access and knowledge.

Changes in institutions and public policies reflect these new economic realities. Administrative reforms have recast government ministries (of machinery, textiles and so on) as industry associations, which now engage in informal discussions and negotiations with official agencies, as do individual companies and interest groups (Kennedy 2005). Fiscal reforms have sought to redress imbalance between central and local revenue shares and to enhance revenue buoyancy to keep pace with growing demands for spending on education, health care, pensions, infrastructure and environment.

Three decades of reform have reshaped China's economy into a hybrid that is increasingly responsive to domestic and international market forces even though some segments, for example, capital markets and investment spending, reflect the continued legacy of planning.

Key Factors in China's Reform Success

Although the period since the late 1970s has brought huge increases in output, productivity, and incomes, China's reforms remain far from complete (Lardy 1998). The costs and inefficiencies associated with unfinished or delayed reform are large. They include remnants of the plan era, for example the underpricing of energy, water, and bank loans, which exacerbates China's environmental and employment problems. Some stem from the reform itself, for instance the continuing epidemic of rent-seeking and graft. Others, including the consequences of weak systems of environmental management, law, public finance, banking, and investment allocation, reflect halfway houses that combine inherited political and economic structures with partial reform efforts (Pei 2006).

How has China's reform achieved so much when its economic system contains so many weak links? China's recent experience encourages us to think of a hierarchy of desirable features that support growth or, if absent, hinder it. These growth enhancing conditions are not equally important. In China, partial measures affecting incentives, prices, mobility, and competition – what we might term 'big reforms' – created a powerful momentum that overwhelmed the friction and drag arising from a host of 'smaller' inefficiencies associated with price distortions, imperfect markets, institutional shortcomings, and other defects that retarded growth and increased its cost but never threatened to stall the ongoing boom (Perkins 1994).

In the presence of large gaps between current and potential output, and of neglected opportunities for expanding the production frontier, limited reform that even partially ruptures the shackles surrounding incentives, marketing, mobility, competition, price flexibility and innovation may accelerate growth. Begin with an economy operating well below its potential, partly because its workers, perceiving that effort hardly affects their incomes, withhold much of their available energy (which itself is reduced by chronic undernutrition). Now restore the link between effort and reward, permit a partial market revival, and open the door to experimentation with international trade and investment. Without disruptive changes in trade flows and political structures that accompanied early reform efforts in the former Soviet Union and Eastern Europe, such simple initiatives – which approximate the circumstances of China's early reforms – can readily ignite a burst of growth, even if prices, financial institutions, judicial enforcement, policy transparency, corporate governance and many other features of the economy remain far from ideal.

A review of what we call 'big reforms' explains the unexpected coincidence of stunning growth with deeply flawed institutions.

Incentives. In China, restoring the link between effort and reward was hugely beneficial even with

large price distortions and a limited market activity. The shift from collective to household farming produced an immediate surge in agricultural production even though the farm sector of the 1980s embodied fewer 'free market' characteristics than Chinese agriculture of the 1920s and 1930s, or even the early 1950s. The same observation applies to private business, which has expanded rapidly and become the largest source of new employment despite its limited access to official support, legal protection and formal credit markets.

Prices. The expansion of price flexibility, most notably through the dual price system, thrust market forces into the economic lives of all Chinese households and businesses. Participants in China's economy – including the large state-owned enterprises at the core of the plan system – suddenly faced a new world in which market prices governed the outcome of marginal decisions to sell above-plan output or to purchase materials and equipment. This partial and gradual liberalization of pricing opened the door to what Naughton (1995) has dubbed 'growing out of the plan,' in which directing incremental output towards market allocation gradually reduced the importance of the plan sector without a political struggle.

Mobility. As the reform progressed, rising urban incomes created new demands for labour in China's cities and towns, especially in construction, services and in new export industries. Responding to this demand, individual villagers began to circumvent regulations that had long barred rural workers from moving to the cities. With the assistance of would-be urban employers and of rural governments, the initial trickle of migration expanded into the largest internal migration in world history.

Partial liberalization of prices, which allowed cash markets to sell food and other necessities with no requirement for residence-based ration tickets, provided essential support for this growing flow of migrant labour. As with the earlier shift from collective to household farming, massive change responded to price signals that, however imprecise, indubitably reflected underlying resource scarcities. Villagers did not need an exact calculation to see that they could raise their incomes by taking up non-farm occupations; several hundred million recognized the opportunity and made the choice.

Competition. Planning attempts to reduce economic uncertainty by pairing suppliers with customers and by specifying the nature of future transactions. Planning also controls the entry of new firms and the exit of weak enterprises. In China, the expansion of incentives, mobility, and markets created unprecedented opportunities to rearrange supply links, to establish new enterprises and to develop existing firms (both domestic and foreign) by commercializing new products and pursuing new markets. Entry squeezed profits (Naughton 1992). The state, as the main owner of enterprise assets, suffered the financial consequences, as the GDP share of fiscal revenue suffered a long decline (Wong and Bird 2008). The resulting fiscal pressures encouraged officials at all levels to respond to pleas from hardpressed enterprises by allowing piecemeal expansion of reform (Jefferson and Rawski 1994).

The scale of entry and exit is startling. The number of industrial firms rose from under 0.4 million in 1980 to nearly 8 million in 1990 and 1996; the 2004 economic census, which excluded enterprises with annual sales below RMB5 million, counted 1.33 million manufacturing firms (Jefferson and Singh 1999, p. 25; Economic Census 2004, pp. 1, 2, 23); in construction, the number jumped from 6,604 to 58,750 between 1980 and 2005, with the latter total excluding subcontractors (Yearbook 2006, p. 579). On the exit side, bankruptcy and restructuring have eliminated many weak firms: between 2001 and 2004, for example, the number of state enterprises in all sectors declined by 177,700 (State Council 2005). Employment in state-owned industry dropped from 45.2 to 8.9 million between 1992 and 2005 (Yearbook 1996, p. 402; 2006, p. 505).

Although Young (2000) and others argue that internal trade barriers limit domestic competition by obstructing the flow of goods and funds across provincial and other administrative boundaries, we believe that the impact of such barriers has faded, allowing rapid expansion of road traffic, telecommunications, chain stores, supply

networks and other new developments to push China's economy towards extraordinarily high levels of competition. Despite pockets of monopoly and episodic local trade barriers, intense competition now pervades everyday economic life. The auto sector provides a perfect illustration: two decades of competition have sucked a lethargic state-run oligopoly into a whirlwind of rivalries in which upstarts such as Chery and Geely wrestle for market share with state-sector heavyweights and global titans. The payoff – rapid expansion of production, quality, variety, and productivity, along with galloping price reductions – has injected a dynamic new sector (not just manufacture of vehicles, components and materials, but also auto dealers, service stations, parking facilities, car racing, publications, motels, tourism, and so on) into China's economy.

The auto sector also illustrates how economic opening has ratcheted up competition throughout China's economy. With few sectors sheltered from imports and with foreign-linked firms participating in a growing array of domestic activities, incumbent suppliers of soybeans, machine tools, retail services, and an endless array of other goods now face competition from rival producers in America, Japan or Brazil as well as Jilin, Zhejiang and Sichuan.

Price wars and advertising, two unmistakable signs of competition, have become commonplace. Chinese newspapers are filled with accounts of fierce price competition among producers of autos, televisions, microwaves, air conditioners, and many other products. Advertising expenditure in 2006 matches total urban retail sales for 1990 (Nielsen Media Research, 2006; Yearbook 2006, p. 678). The decline of former industry leaders like Panda (televisions) and Kelon (home appliances) and the ascent of new pacesetters like Wahaha (beverages), Wanxiang (auto parts) and Haier (home appliances) from obscure beginnings show how competition has added new fluidity to Chinese market structures.

Innovation. Prior to reform, China experienced a general failure of dynamic efficiency. Under the plan system, apart from exceptional instances of direct highlevel intervention ('innovation by order'), producers neglected innovation in favour of pursuing short-term targets for physical output ('fulfilling the plan'). As a result, the expansion of society's production frontier lagged behind the potential embodied in available knowledge and resources. The consequences are readily visible: First Auto Works, one of China's premier manufacturers, found its 'obsolescence of equipment and models worsening day by day' following '30 years of standing still' under the planned economy (Li Hong 1993, p. 83).

Reform put an end to this stand-pat mentality by widening the gap between financial outcomes for strong and weak firms, their managers and their employees. The presence of price distortions, subsidies and official intervention could not obscure the central issue: do we pursue innovation in order to maintain and perhaps expand our sales, market share, profits, wages, and employment security, or do we sit tight and hope that current or potential rivals do not leave us behind? Especially since China's entry into WTO, the proportion of firms engaging in R&D has grown rapidly, as has the ratio of R&D spending to GDP (Hu and Jefferson 2008).

On the supply side, efforts to upgrade the quality and variety of products benefited from rapid increases in China's supply of educated workers. China's growing engagement with the global economy created immense inflows of new technology, not just from imports of equipment and know-how, but from new links connecting millions of Chinese workers, engineers, and managers with the technical standards, engineering processes and management practices needed to compete in global markets.

Key Elements in the Political Economy of Chinese Reform

What of the policy process associated with these extraordinary changes? Despite the authoritarian nature of China's political system, pre-reform policy structures allowed widespread experimentation and regional variation within broad guidelines set at the centre. This encouraged local officials to develop strategies whose success

might attract high-level attention and also allowed national leaders to 'play to the provinces' (Shirk 1993) by assembling coalitions of like-minded officials to demonstrate the merits of their preferred policy options and to lobby for nationwide implementation of those policies.

This arrangement, under which national policies emphasized broad principles or parameters rather than specific instructions or regulations, continued into the reform period. What changed is the content of the directives articulated at the centre, formerly directed towards ideological matters, which now focused increasingly on issues surrounding economic growth.

Looking beyond the principles emanating from the top, we see three additional elements as completing the skeleton of China's reformist political economy. *Decentralization* endows provinces and localities with both the resources and the incentive to experiment with local approaches to specific policies (for example, rural industrialization) and difficulties (for example how to deal with redundant statesector workers), providing they observe central guidelines. *Competition* within the political system is not new, but now focuses on economic outcomes, which exercise increasing leverage over the career paths of leaders at every level. Continued promotion and recruitment of leaders whose reputation and career prospects rest on past and future economic success has gradually created a large and expanding *coalition among growth-minded, market oriented individuals and groups* within China's policy elite, whose power and influence helps to shift the content of central guidelines towards market outcomes.

Broad Guidelines: What They Can and Cannot Do

Chinese tradition emphasizes the government of men (and, beginning in the late 20th century, some women) rather than laws. In the absence of detailed instructions, how do China's top leaders direct the behaviour of lower-level governments and individual officials? Functionaries at all levels study and discuss the speeches and writings of top leaders, which lay out the desired course of public policy and explain what lower levels of officialdom should and should not do. These guidelines become encapsulated in catchy slogans that gain wide currency. In turn, these slogans, and the policy guidelines that inform them, direct the flow of policy implementation at all levels.

From the start of China's reform in the late 1970s, these directives increasingly emphasized economic matters. Indeed, China's political economy has come to rest on a grand but unspoken bargain between the Communist Party and the Chinese public in which the party ensures economic growth and promotes China's global standing in return for public acquiescence to its autocratic rule and anachronistic ideology (Keller and Rawski 2007b). As a result, the articulation and fulfilment of key economic objectives now constitute core ingredients in extending the political legitimacy of the Chinese state. Economic objectives embedded in documents, speeches, and slogans reverberate at every level of society, where they become benchmarks for evaluating current or proposed actions. Deng Xiaoping's praise of reform during his southern tour of 1992 was widely seen as a favourable signal for policy innovations, including many that received no specific mention from him. In similar fashion, emphasis (or omission) of praise for 'small and medium enterprises' will be interpreted as high-level encouragement of (or caution against) policies favouring private business.

Decentralized Experimentation

The experience of the 20th century surely qualifies the Chinese as the world's leading practitioners of economic experimentation. China's reform economy amply displays this characteristic. We see the national government conducting trials of novel institutions, for example 'special economic zones', while provinces and localities develop their own variations of pension systems, industrial regulation, and so on.

The decentralization of industry, which placed all but the largest enterprises under the control of

lower-level governments, and of public finance, which, especially prior to the 1994 fiscal reforms, assigned major revenue streams to provincial and local administrations, provided regional and local governments with ample resources with which to pursue such experimentation.

Competition

Prior to the inception of reform, China developed a tradition of policy entrepreneurship in which local figures compete for high-level attention by demonstrating the beneficial implementation of the principles enshrined in broad central directives. This competition intensified under the reform, with GDP growth and other economic criteria replacing ideological benchmarks as the arbiters of success. Thus Li and Zhou (2005) find that promotion prospects for provincial leaders rise, and the likelihood of termination declines as provincial economic performance improves. Whiting (2001) makes similar observations about local officials.

Officials at all levels possess the authority as well as the resources needed to promote local growth. They also have strong incentives to do so, because their career prospects, as well as personal financial opportunities for themselves and their families, are closely tied to the economic trajectory of the jurisdictions under their leadership. Growth expands the pools of public revenue and enterprise profits over which officials exercise varying degrees of control, enlarges business opportunities available to the families and associates of local leaders, and swells the flow of (legal and illicit) rents directed towards official agencies and their managers.

These circumstances have transformed China's local and provincial governments into eager champions of development, each striving to outdo its neighbours in expanding infrastructure and strengthening the foundations of 'pillar industries'. This competition contributes mightily to the persistent 'investment hunger' visible in China's economy, as local administrations resist central calls for restraint in enlarging existing facilities and building new ones.

Pro-growth Coalition

China's reform leaders, like politicians everywhere, endeavour to appoint and promote like-minded successors and subordinates. As Shirk (1993) and others have noted, the reform movement's initial successes acted as a powerful recruiting device, with the lure of rich payoffs adding many influential converts to the cause of reform. As the reform gained momentum, the circulation of elites, including the assignment of successful officials to lagging regions for the express purpose of jump-starting growth, created mentor-student relationships between growth-oriented officials and increasing numbers of would-be imitators. The widespread practice of sending study teams to absorb the 'advanced experiences' of dynamic localities further expanded the reform constituency among China's policy elites.

Of particular importance is the legacy of the Cultural Revolution, which truncated educational opportunities for whole cohorts of Chinese. This historical accident created a unique opportunity to advance the reform agenda. When the retirement of Deng Xiaoping and other 'revolutionary elders' focused attention on generational change, reformist leaders managed to bypass the customary emphasis on seniority, skipping over the 'lost generation' of Cultural Revolution victims to promote younger candidates. The increasing prominence of university graduates, including returnees from overseas study and young professionals with close ties to international business, accelerated the development of what became a loose and unorganized but increasingly potent coalition of like-minded officials whose objectives centred on growth-promoting and increasingly market-oriented reforms.

Despite these gains, the evolution of policy towards private business demonstrates the difficulty of translating power and influence into genuine institutional change. Legal documents confirm the painfully slow expansion of official protection. At the start of reform, private business operated in a legal limbo. Some entrepreneurs disguised their firms as collectives; others purchased informal protection from powerful

individuals or agencies. A succession of amendments to China's 1982 constitution slowly expanded recognition of the non-public economy, first as a 'complement' to the state sector (1988), than as an 'important component' (1999) of the 'socialist market economy' (itself a new term dating from 1993). The 'Law on Solely Funded Enterprises', which took effect in 2000, guaranteed state protection for the 'legitimate property' of such firms, but without using the term 'private' or specifying any agency or process to implement this promise.

Further constitutional amendments adopted in 2004 breached the former taboo on the term 'private' by stating that 'citizens' lawful private property is inviolable'. The long march towards official recognition of private business came to an end only in 2007 when, following five years of fierce debate, China's legislature enacted a landmark Property Rights Law which, for the first time, explicitly places privately held assets on an equal footing with state and collective property.

Conclusion

Reform has delivered enormous economic gains despite deep and potentially dangerous flaws in China's institutions and policy structures. The same framework of structures and incentives that spurs rapid economic advance also generates ambiguous and often disturbing consequences along other socioeconomic dimensions. Environment and inequality illustrate the range of outcomes.

Economy (2004) and others demonstrate how China's unbridled rush to maximize GDP growth, together with weak regulatory and legal structures, has produced environmental degradation on a scale that far exceeds internationally acceptable standards. Historical comparisons also show that improved technology and the spread of environmental consciousness among China's growing middle class are pushing China towards regulation and remediation of atmospheric and water pollution at an earlier stage of the development process than occurred in Japan, Korea, or the United States.

China's reforms have literally pulled hundreds of millions out of poverty, especially in the countryside. Reform has also increased China's income inequality to levels that now approach some of the highest in the developing world. Although attention focuses on income gaps between urban and rural areas and between coastal and interior provinces, growing income differences between neighbours within provinces and within the urban and rural sectors account for most of the increase in inequality (Benjamin et al. 2008). In rural areas, this increase is tied to the disequalizing role of some forms of non-agricultural income, and laggard growth of farming income, especially beginning in the mid-1990s. In urban areas, a decline in the role of subsidies and entitlements, increasing wage inequality related to labour market and enterprise reform, and the effect of SOE restructuring on some cohorts and households have enlarged the dispersion of incomes. Rising returns to human capital and differences in access to education have widened income differences in all sectors. Corruption, although difficult to quantify, may also have contributed to growing inequality of wealth and welfare.

Despite these and other difficulties, China's recent experience demonstrates that activating key economic drivers, including incentives, mobility, prices, competition, and innovation, can unleash sufficient momentum to overwhelm a variety of system costs. China's economic boom, in 2007 completing its third decade, rests on a unique set of historical circumstances, some favourable, others less so.

China's success cannot ensure the efficacy of 'Chinese policies' in other times and places. There is also no guarantee that the mechanism described in this article can enable China to extend its enviable record of high speed growth. Even so, China's continuing accumulation of physical resources and human capital, the intense focus of public policy on promoting growth, and the willingness of China's leaders to implement bold initiatives create a favourable climate for further reform and continued economic expansion.

See Also

- China, Economics in
- Dual Track Liberalization
- Maoist Economics
- Soft Budget Constraint

Bibliography

Benjamin, D., L. Brandt, J. Giles, and S. Wang. 2008. Income inequality during China's economic transition. In Brandt and Rawski (2008).

Brandt, L., and T.G. Rawski, eds. 2008. *China's great economic transformation*. New York: Cambridge University Press.

Brandt, L., T.G. Rawski, and X. Zhu. 2007. International dimensions of China's long boom: Trends, prospects and implications. In *China and the balance of influence in Asia*, ed. W.W. Keller and T.G. Rawski. Pittsburgh: University of Pittsburgh Press.

Brandt, L., C.-T. Hsieh, and X. Zhu. 2008. Growth and structural transformation in China. In Brandt and Rawski (2008).

Economic Census. 2004. *Zhongguo jingji pucha nianjian 2004* [Yearbook of China's 2004 economic census]. Vol. 4. Beijing: Zhongguo tongji chubanshe.

Economy, E. 2004. *The River runs black: The environmental challenge to China's future*. Ithaca: Cornell University Press.

Heston, A., R. Summers, and B. Aten. 2006. Penn World Table Version 6.2, Center for International Comparisons of Production, Income and Prices at the University of Pennsylvania, September. Online. Available at http://pwt.econ.upenn.edu. Accessed 15 June 2007.

Hu, A.G.Z., and G.H. Jefferson. 2008. Science and technology in China. In Brandt and Rawski (2008).

Jefferson, G.H., and T.G. Rawski. 1994. Enterprise reform in Chinese industry. *Journal of Economic Perspectives* 8(2): 47–70.

Jefferson, G.H., and I. Singh. 1999. *Enterprise reform in China: Ownership, transition, and performance*. Oxford/New York: Oxford University Press.

Keller, W.W., and T.G. Rawski. 2007a. *China and the balance of influence in Asia*. Pittsburgh: University of Pittsburgh Press.

Keller, W.W., and T.G. Rawski. 2007b. China's peaceful rise: Roadmap or fantasy?. In Keller and Rawski.

Kennedy, S. 2005. *The business of lobbying in China*. Cambridge: Harvard University Press.

Lardy, N.R. 1983. *Agriculture in China's modern economic development*. Cambridge: Cambridge University Press.

Lardy, N.R. 1998. *China's unfinished economic revolution*. Washington, DC: Brookings Institution.

Lardy, N. 2002. *Integrating China into the global economy*. Washington, DC: Brookings Institution.

Lardy, N.R., and K. Lieberthal. 1983. *Ch'en Yün's strategy for China's development: A non-maoist alternative*. Armonk: M.E. Sharpe.

Li, H. 1993. *Zhongguo qiche gongye jingji fenxi* [Economic analysis of China's auto industry]. Beijing: Zhongguo renmin daxue chubanshe.

Li, X. 2006. *Assessing the extent of China's marketization*. Aldershot: Ashgate.

Li, H., and L.-A. Zhou. 2005. Political turnover and economic performance: The incentive role of personnel control in China. *Journal of Public Economics* 89: 1743–1762.

Lin, J.Y. 1992. Rural reforms and agricultural growth in China. *American Economic Review* 82: 34–51.

National Bureau of Statistics. 2006. National bureau of statistics of China. Statistical Communiqué of the People's Republic of China on the 2006 National Economic and Social Development. Beijing: China Statistics Press.

Naughton, B. 1992. Implications of the state monopoly over industry and its relaxation. *Modern China* 18: 14–41.

Naughton, B. 1995. *Growing out of the plan: Chinese economic reform, 1978–1993*. Cambridge/New York: Cambridge University Press

Nielsen Media Research. 2006. Review of the Chinese advertising market in 2006.

OECD (Organisation for Economic Co-operation and Development). 2005. *OECD economic surveys: China*. Paris: OECD.

Pei, M. 2006. *China's trapped transition: The limits of developmental autocracy. Cambridge*: Harvard University Press.

Perkins, D.H. 1994. Completing China's move to the market. *Journal of Economic Perspectives* 8(2): 21–46.

Perkins, D.H., and T.G. Rawski. 2008. Forecasting China's economic growth to 2025. In Brandt and Rawski (2008).

Shirk, S.L. 1993. *The political logic of economic reform in China*. Berkeley: University of California Press.

Sicular, T. 1995. Redefining state, plan, and market: China's reforms in agricultural commerce. *China Quarterly* 144: 1020–1046.

State Council. 2005. Report No. 1 of the main data from the first national economic census. Chinese document from the Office of the State Council Leading Small Group for the First National Economic Census and the National Bureau of Statistics, 6 Dec 2005.

Whiting, S.H. 2001. *Power and wealth in rural China: The political economy of institutional change*. Cambridge: Cambridge University Press.

Wong, C.P.W., and R. Bird. 2008. China's fiscal system: A work in progress. In Brandt and Rawski (2008).
Yearbook. 1996. *Zhongguo tongji nianjian 1996*. [China statistical yearbook 2006]. Beijing: China Statistics Press.
Yearbook. 2006. *Zhongguo tongji nianjian 2006*. [China statistical yearbook 1996]. Beijing: China Statistics Press.
Young, A. 2000. The razor's edge: Distortions and incremental reform in the People's Republic of China. *Quarterly Journal of Economics* 115: 1091–1136.

Choice of Technique and the Rate of Profit

N. Okishio

Capitalistic Criterion

In a capitalistic economy the main production decisions are made by private capitalists. The choice of technique is one of the decisions in their hands, and the criterion for that choice is to maximize the expected profit rate. In order to calculate that rate they must have expectations of the prices of various commodities, and of the wage rate.

Assuming a linear technology, in the *i*th sector capitalists have T_i alternative techniques:

$$a_{i1}(k_i), a_{i2}(k_i), \ldots, a_{in}(k_i), \tau_i(k_i)$$
$$k_i = 1, 2, \ldots, T_i$$

where $a_{ij}(K_i)$ is the amount of the *j*th commodity used as input to produce one unit of the *i*th commodity by the k_ith technique and $\tau_i(K_i)$ is the amount of labour necessary to produce one unit of the *i*th commodity by the k_ith technique.

Capitalists have expected prices and the wage rate:

$$p_1^e, p_2^e, \ldots, p_n^e, w^e,$$

where p_i^e is the expected price of the *i*th commodity and w^e is the expected wage rate.

The expected profit rate from the k_i technique, which is denoted as $r_i^e(k_i)$, is calculated as

$$p_i^e = \left[1 + r_i^e(k_i)\right]\left[\sum a_{ij}(k_i)p_i^e + \tau_i(k_i)w^e\right].$$

Capitalists choose the technique which yields the highest expected profit rate. If

$$r_i^e\left(k_i^e\right) \geqq r_i^e(k_i) \quad k_i = 1, 2, \ldots, T_i \tag{1}$$

then they choose the k_i^* th technique among T_i alternatives. As is easily seen, (1) can be rewritten as

$$\sum a_{ij}\left(k_i^*\right)p_i^e + \tau_i\left(k_i^*\right)w^e \leqq \sum a_{ij}(k_i)p_j^e + \tau_i(k_i)w^e$$
$$k_i = 1, 2, \ldots, T_i \tag{2}$$

The means that the expected unit cost is smallest in the k_i^* th technique. So in this case *the maximum profit rate criterion* is equivalent to *the minimum unit cost criterion*. However, this equivalence does not hold in general. If we introduce durable equipment the two criteria are not equivalent. But for simplicity here we will ignore durable equipment.

Profit Rate and Techniques

In the *i*th sector by the minimum unit cost criterion capitalists adopt the technique

$$a_{i1}, a_{i2}, \ldots, a_{in}, \tau_i$$

and labourers receive the commodity basket

$$b_1, b_2, \ldots, b_n$$

per unit of labour. Then an equal rate of profit *r* between *n* sectors is determined by the following equations:

$$p_i = (1 + r)\left(\sum_{j=1}^{n} a_{ij}p_j + \tau_i w\right)$$
$$i = 1, 2, \ldots, n \tag{3}$$
$$w = \sum_{i=1}^{n} b_i p_i$$

From these equations it is clear that the profit rate r depends on techniques (a_{ij}, τ_i) and the real wage basket (b_i).

In order to examine the relationship between the profit rate and techniques in various sectors, we must introduce a new concept: *basic sectors*. P. Sraffa has used this terminology, defining basic sectors as those whose outputs are directly or indirectly necessary in the production of every commodity (see ch. 2 in Sraffa 1960).

However, it is not guaranteed a priori that such basic sectors exist; and even if they do exist, the concept is not useful for our purpose here.

Now we redefine basic sectors as those whose products are wage goods, or whose products are directly or indirectly necessary to produce wage goods. 'Wage goods' means commodities which are included in the real wage basket (b_1,. . .,b_n). If $b_i > 0$ then the ith commodity is a wage good. Basic sectors in this sense necessarily exist, for there must be at least one commodity which is a wage good.

Suppose there are m basic sectors, with $m \leqq n$; after renumbering, let the 1st, 2nd,. . .,mth sectors be basic sectors. Then the equations

$$\begin{aligned} p_i &= (1+r)\left(\sum_{j=1}^{m} a_{ij}p_j + \tau_i w\right) \\ & i = 1, 2, \ldots, m \\ w &= \sum_{i=1}^{m} b_i p_i \end{aligned} \qquad (4)$$

are sufficient to determine the profit rate r, where prices (p_1,. . .,p_m) and the wage rate w are both positive. Therefore we can say that the profit rate does not depend on techniques in the non-basic sectors. For example, pure luxury goods are non-basic commodities. Whatever great improvement may occur in the techniques in those sectors, the (equalled) rate of profit is not influenced at all. This conclusion was first found by Ricardo, but Marx did not accept it (Ricardo 1821, p. 132; Marx 1867–94, Vol. III, ch. 5, pp. 83–4).

Hereafter in this essay we confine ourselves to techniques in the basic sectors only, and we assume that the classification into basics and non-basics remains unaffected by the technical changes considered here.

Technical Progress

Let us suppose the profit rate r to be determined by Eq. 4, and that in the kth sector ($1 \leqq k \leqq m$) a *new* alternative technique

$$a'_{k_1}, a'_{k_2}, \ldots, a'_{k_m}, \tau'_k$$

becomes feasible. Capitalists must then calculate the expected profit rate of this new technique and compare it with those of alternative techniques to decide whether or not to adopt it, so we now need an assumption about how capitalists form their expectations. For simplicity we assume

$$p_i^e = p_i, \; w^e = w \qquad i = 1, 2, \ldots, m$$

i.e. capitalists expect that current prices and the wage rate, as given by Eq. 4, will remain the same (static expectations).

If the following inequality holds, capitalists adopt the new technique:

$$\sum_{j=1}^{m} a'_{kj}p_j + \tau'_k w < \sum_{j=1}^{m} a_{kj}p_j + \tau_k w \qquad (5)$$

Supposing this to be so, the previous technique in the kth sector (a_{k1}, $a_{k\,2}$,. . ., a_{km},τ_k) is replaced by the new technique $\left(a'_{k_1}, a'_{k_2}, \ldots, a'_{k_m}, \tau'_k\right)$. How does the profit rate r as given by Eq. 4 then change, under the requirement that the real wage basket remain unchanged? We can prove that the profit rate r necessarily rises, as follows:

Putting

$$\beta = 1/(1+r), \qquad q_i = p_i/w,$$

Eq. 4 are rewritten as

$$\beta q_i = \sum a_{ij} q_j + \tau_i \qquad i = 1, 2, \ldots, m \qquad (6)$$

$$1 = \sum b_i q_i \qquad (7)$$

Let the solution of Eqs. 6 and 7 be

$$(\beta, q_1, \ldots, q_m)$$

When $(a_{k1}, a_{k\ 2}, \ldots, a_{km}\tau_k)$ is replaced by $\left(a'_{k_1}, a'_{k_2}, \ldots, a'_{k_m}, \tau'_k\right)$, the profit rate is determined by

$$\beta q_i = \sum a_{ij} q_j + \tau_i,\ i = 1, \ldots, k-1, k+1, \ldots, m \tag{8}$$

$$\beta q_k = \sum a'_{kj} q_j + \tau'_k \tag{9}$$

and Eq. 7. Let the solution of Eqs. 8, 9 and 7 be

$$(\beta', q'_1, \ldots, q'_m).$$

As $q'_i > 0$ for all i, the coefficients matrix of q_i ($i = 1,2,\ldots,m$) satisfies the Hawkins-Simon conditions (see Simon and Hawkins 1949).

From Eq. 6 to Eq. 9, we get

$$\beta' \Delta q_i = \sum a_{ij} \Delta q_j - q_j \Delta\beta,\quad i = 1, \ldots, k-1, k+1, \ldots, m \tag{10}$$

$$\beta' \Delta q_k = \sum a'_{ij} \Delta q_j - q_k \Delta\beta + \left\{\sum q_j \Delta a_{kj} + \Delta\tau_k\right\} \tag{11}$$

$$0 = \sum b_i \Delta q_i \tag{12}$$

where

$$\Delta q_i = q'_i - q_i, \quad \Delta\beta = \beta' - \beta$$
$$\Delta a_{kj} = a'_{kj} - a_{kj}, \quad \Delta\tau_k = \tau'_k - \tau_k$$

The third term on the right side of Eq. 11 is negative, by Eq. 5. If $\Delta\beta \geqq 0$, then in Eqs. 10 and 11 $\Delta q_k < 0$ and $\Delta q_i \leqq 0$ for all $i \neq k$, because as shown above the coefficient matrix of Δq_i in Eqs. 10 and 11 satisfies the Hawkins–Simon conditions. If the kth commodity is a wage good, $\Delta q_i < 0$ contradicts Eq. 12. If the kth commodity is a means of production (that is it belongs to the basic sectors), there must be at least one kind of wage good whose $\Delta q_i < 0$; again this contradicts Eq. 12. So $\Delta\beta > 0$, or in other words the profit rate r rises.

The proposition that any new technique which satisfies the profit rate criterion Eq. 5 and so is introduced into the basic industries necessarily increases the general rate of profit, cannot be compatible with the Marxian law of the tendency for the profit rate to fall. However large the organic composition of production may become, the general rate of profit must increase without exception, provided that the newly introduced technique satisfies the profit rate criterion and the rate of real wage remains constant (see Okishio 1961, for further discussion).

Joint Production

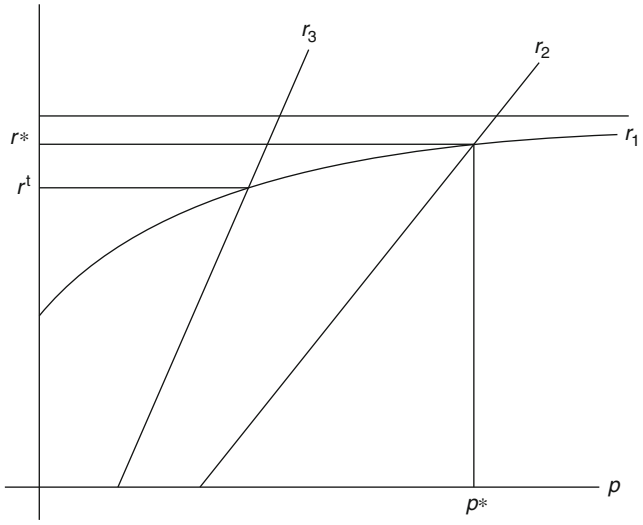
So far we have disregarded joint production as well as durable equipment. Even if we introduce durable equipment, the conclusion obtained in the former section still holds (see Nakatani 1984). However, when we consider the joint production it is possible (though not necessary) to find a case in which the proposition does not hold, a perverse conclusion that was originally presented by Salvadori (1981).

In order to show such a case we examine the following numerical example. In this economy, shown in Table 1, there are two kinds of commodity. The second commodity is a wage good and it is produced jointly with the first commodity. Let

Choice of Technique and the Rate of Profit, Table 1

Input			Output		
Technique	1	2	Labour	1	2
1	0.5	0.5	1	1	2
2		0.5	1	1	
3		0.5	1	2	

Choice of Technique and the Rate of Profit, Fig. 1



the real wage rate be 0.7 unit of the second commodity. At the first stage we assume that techniques 1 and 2 only are feasible.

The profit rates of techniques 1 and 2 are determined by

$$p_1 + 2p_2 = (1 + r_1)(0.5p_1 + 0.5p_2 + w)$$
$$p_1 = (1 + r_2)(0.5p_2 + w)$$
$$w = 0.7p_2$$

where r_1, r_2 are the profit rates of techniques 1 and 2, respectively. Putting $p = p_1/p_2$ these equations are rewritten as

$$p + 2 = (1 + r_1)(0.5p + 1.2)$$
$$p = 1.2(1 + r_2)$$

The profit rates of both technique are drawn on Fig. 1.

Now we examine the condition in which both techniques 1 and 2 are used. If technique 1 only is used at activity level x, then the surplus products consist of $(1–0.5)x$ units of commodity 1 and $(2–0.5 \simeq 0.5)x$ units of commodity 2. Therefore if the capitalists' demand for the surplus products (for their consumption or investment) are 100 units of commodity 1 and 50 units of commodity 2, then there must be excess demand for commodity 1 or excess supply for commodity 2 and the relative price p_1/p_2 increases.

When p rises above p^* the expected profit rate r_2 becomes greater than r_1, so technique 2 is introduced. However, technique 2 cannot replace technique 1 completely because technique 2 cannot produce commodity 2. Therefore both techniques must be used, which requires that the equal rate of profit be determined at r^*.

At the next stage we assume that technique 3 becomes feasible. Technique 3 is apparently superior to technique 2, because in the new technique capitalists get more output from the same input, so it replaces technique 2. The profit rate of technique 3 is calculated from

$$2p_1 = (1 + r_3)(0.5p_2 + w)$$
$$w = 0.7p_2$$

which can be rewritten as

$$2p = 1.2(1 + r_3)$$

This equation is also plotted on Fig. 1, from which it can be seen that the equalized rate of profit falls to r'.

Substitutional Technical Change

In the previous sections we treated the relationship between the profit rate and technical change

under the condition that the real wage basket remain unchanged. The change in the technique adopted was not induced by a change in the real wage rate, but was caused by the introduction of a new process.

The question now is the relationship between the profit rate and technical change that is induced by a change in the real wage rate. We define the level of the real wage rate λ as follows.

Assume that each labourer spends his wage income on various wage goods in fixed proportions. Then we have

$$w = \lambda \sum b_i p_i,$$

where the b_i are all constant and λ is the level of the real wage rate.

At the first stage $\lambda = 1$, and the profit rate is determined by Eqs. 6 and 7. At the next stage $\lambda > 1$, which means a rise in the real wage rate. Then the profit rate is determined by the following equations

$$\beta q_i = \sum a'_{ij} q_j + \tau'_i \qquad i = 1, 2, \ldots, m \quad (13)$$

$$1 = \lambda \sum b_i q_i \quad \lambda > 1. \quad (14)$$

As shown in Eq. 13 the techniques used in every sector may differ from the techniques used at the first stage, because of the rise of the real wage rate and the change of prices which accompanies it. What can we say about the relationship between the newly adopted technique a'_{ij}, τ'_i and the old technique (a_{ij}, τ_i) ? (Of course we assume that no technique becomes newly feasible for capitalists between the first stage and the next stage.)

Let the solution of Eqs. 6 and 7 be

$$(\beta, q_i, \ldots, q_m).$$

Then

$$\sum a_{ij} q_j + \tau_i \leqq \sum a'_{ij} q_i + \tau'_i \quad (15)$$

because technique (a_{ij}, τ_i) would not have been adopted at the first stage if inequality Eq. 15 had *ij i*not held; rather they would have adopted (a'_{ij}, τ'_i).

Let the solution of Eqs. 13 and 14 be

$$(\overline{\beta}, \overline{q}_1, \ldots, \overline{q}_m).$$

Then, arguing as for Eq. 15,

$$\sum a_{ij} \overline{q}_j + \tau_i \geqq \sum a'_{ij} \overline{q}_j + \tau'_i. \quad (16)$$

Using inequalities Eqs. 15 and 16 we can prove that in going from the first stage to the second the profit rate necessarily falls, as follows:

From Eqs. 6, 7, 13 and 14 we get

$$\overline{\beta} \delta q_i = \sum a'_{ij} \delta q_j + q_i \delta \beta + \left\{ \sum_{i = 1, 2, \ldots, m} \left(a'_{ij} - a_{ij} \right) q_j + \left(\tau'_i - \tau_i \right) \right\} \quad (17)$$

$$0 = \sum b_i \delta q_i + (\lambda - 1) \sum b_i \overline{q}_i \quad (18)$$

where

$$\delta q_i = \overline{q}_i - q_i, \qquad \delta \beta = \overline{\beta} - \beta.$$

From Eq. 15 we know that the third term on the r.h.s. of Eq. 17 is non-negative. If we assume $\delta\beta \leqq 0$ then all the δq_i become non-negative because the coefficient matrix of δq_i satisfies the Hawkins-Simon conditions. But since $\lambda > 1$ that contradicts Eq. 18. So $\delta\beta$ must be positive, or in other words the profit rate must fall.

When the real wage rate rises capitalists cannot avoid a fall in the profit rate, even if they substitute techniques to avoid it; only the introduction of new and superior feasible techniques can prevent the fall. However, we cannot say that the capitalists' efforts to substitute with exciting techniques are of no use to them. Though they cannot avoid a fall in the profit rate, they can mitigate it. We can prove this as follows.

If in spite of the rise in the real wage rate capitalists adhere to the techniques adopted at the first stage, the profit rate is determined by

$$\beta q_i = \sum a_{ij} q_j + \tau_i \qquad i = 1, 2, \ldots, m \quad (19)$$

$$1 = \lambda \sum b_i q_i \quad \lambda > 1 \tag{20}$$

Let the solution of Eqs. 19 and 20 be

$$\left(\beta^*, q_1^*, \ldots, q_m^*\right).$$

From Eqs. 13, 14, 19 and 20 we get

$$\begin{aligned} \beta^* dq_i = & \sum a_{ij} dq_j - q_i^* d\beta \\ & + \left\{ \sum \left(a'_{ij} - a_{ij}\right) \overline{q}_j + \left(\tau'_i - \tau_i\right) \right\} \\ & i = 1, 2, \ldots, m \end{aligned} \tag{21}$$

$$0 = \lambda \sum b_i dq_i \tag{22}$$

where

$$dq_i = \overline{q}_i - q_i^*, \qquad d\beta = \overline{\beta} - \beta^*.$$

From Eq. 16 the third term of the r.h.s. of Eq. 21 is non-positive. However, if we now consider the case in which substitution actually occurs, then for some i the third term on the r.h.s. of Eq. 21 is negative. If we assume $d\beta \geqq 0$, then all the dq_i become non-positive and some actually negative, because the coefficient matrix of dq_i satisfies the Hawkins–Simon conditions. This contradicts Eq. 22 so $d\beta < 0$. In other words, when the substitution is carried out the profit rate is greater than it would have been if capitalists had adhered to the old optimal technique, which corresponded with the former level of the real wage rate.

See Also

- ▶ Investment Decision Criteria
- ▶ Investment Planning
- ▶ Non-substitution Theorems

Bibliography

Marx, K. 1867–94. *Capital*. Translated from the third German edition by Samuel Moore and Edward Aveling, ed. Frederick Engels. Reprinted. New York: International Publishers, 1967.

Nakatani, T. 1984. Technical change and the rate of profit: Considering fixed capital. *Kobe University Economic Review* 30: 65–78.

Okishio, N. 1961. Technical changes and the rate of profit. *Kobe University Economic Review* 7: 85–99.

Ricardo, D. 1821. On the principles of political economy and taxation. Vol. 1, of The works and correspondence of David Ricardo, ed. P. Sraffa. Cambridge: Cambridge University Press, 1951.

Salvadori, N. 1981. Falling rate of profit with a constant real wage: An example. *Cambridge Journal of Economics* 5(1): 59–66.

Simon, H.A., and D. Hawkins. 1949. Some conditions of macro-economic stability. *Econometrica* 17: 245–248.

Sraffa, P. 1960. *Production of commodities by means of commodities: Prelude to a critique of economic theory*. Cambridge: Cambridge University Press.

C

Chrematistics

M. I. Finley

A Greek word occasionally taken over into English (and some other western European languages) to mean ‘money-getting’, often but not always with pejorative overtones. After a great flurry in the fifth and fourth centuries BC in the original Greek, the word became uncommon and would not be worth noticing here were it not that the major debates among Greek thinkers, chiefly ethical, were revived in the thirteenth century by the scholastic philosophers though without actually using the term ‘chrematistic’. The earliest English example given by the *Oxford English Dictionary* dates from the mid-eighteenth century, in Henry Fielding’s last novel, *Amelia* (1752): ‘I am not the least versed in the chrematistic art . . . . I know not how to get a shilling, or how to keep it in my pocket if I had it’ (Book IX, ch. 5). This pedantry implied no familiarity by readers, for Fielding promptly gave them the sense of it. Nor is there much to be drawn from the unimportant nineteenth-century terminological disagreement over the propriety of the term ‘political economy’, for which Gladstone and a few others preferred ‘chrematistics’.

The word had a long and complex history in early Greek, which we cannot trace properly because of insufficient evidence. The ultimate root was the verb *chrao*, to 'need' or 'use', hence *chrema* (more common in the plural form, *chremata*), 'goods', 'property', 'wealth', and the verbal forms meaning 'to seek wealth' (Other extensions, such as 'to engage in discussion or negotiations' or 'to consult an oracle' need not concern us.). Alongside these *chrao* words there were others with the same sense of goods, property, wealth, but all attempts to codify distinctions in usage, for instance between real property and moveable wealth, have proved not to reflect the actual practice of Greek speakers and writers.

Concomitant with this linguistic history there was a related development in commercial practice and attitudes. One thread that persisted in attitudes, however, was a distrust, even outright condemnation, of trading for profit, of making a profit out of the sheer act of exchange without any additions to, or transformation of, the goods being traded. This motif was already apparent in the *Odyssey*: when Odysseus declined the invitation to join in the games arranged following a feast at the court of King Alcinous of the Phaeacians, he was taunted by one of the courtiers with an unbearable insult:

> No indeed, stranger, I do not think you are like a man of games, . . . but like one who travels with a many-benched ship, a master of sailors who traffic, one who remembers the cargo and is in charge of merchandise and coveted gains (*Odyssey*, 8, 145–64).

This distrust of trade can be amply exemplified right through pagan antiquity and then among the Church Fathers (on whom see Baldwin 1959, pp. 12–16), with differences in nuance from culture to culture and from author to author that are sometimes interesting and at times paradoxical in their practical implications. Cicero provides a neat illustration. In the *De officiis* (1, 150–1) he dismisses those 'who buy from merchants in order to re-sell immediately, for they would make no profit without much outright lying', whereas commerce that is 'large-scale and extensive, importing much from all over the distributing to many without misrepresentation (*vanitas*), is not to be greatly censured'. In that nuance there was an exception because of the social usefulness of the large-scale merchant in his role as importer, with the consequent suggestion that he, unlike the petty trader in the market, was able to avoid outright lying. Not all moralists allowed such an exception, and Cicero gives no hint why one was in practice possible in this particular instance.

What I have been calling a paradox revealed itself when the moral judgements of a given culture or society conflicted with its legal system and rules of practice. Obviously large numbers of Greeks and Romans behaved contrary to the norms of commercial exchange laid down by Aristotle or Cicero, and on the whole both Greek and Roman law accepted the validity of private agreements provided only that they did not require actions specifically prohibited by the law. However, there were also major differences between the two, most obviously with respect to usury. Barring unimportant exceptions, especially in emergency situations, it was the rule among Greek states that no attempt was made to restrict or otherwise regulate rates of interest, whereas from earliest times the Roman law kept a tight reign on moneylending. Hence in Rome the professional moneylender, the *fenerator*, was a figure of distrust and contempt while men of means ranked moneylending at legal rates second only to land ownership as a source of income, and did not often blanch at charging exorbitant rates for loans to communities outside the sphere of the Roman law. There we see several paradoxes at work simultaneously.

This apparent detour has brought us back to the heart of the discussion of money-getting, of chrematistics. We can examine the formulations of moralists and we can match them against the legal precepts, but the critical question of practice in actual exchanges largely escapes us. Not only are we driven to outright guesses about pricing procedures (as about the actual practice with respect to interest charges), but there is no useful ancient testimony about the mechanism of price determination. Much of what is written about the meaning or the reality behind Aristotle's ambiguous formulation, or about Aquinas' concept of just

price, is nothing but a backward projection of a modern author's own notions, for which there is no warrant (and often contrary evidence) in the ancient texts themselves (Finley 1970).

In the end, we have only one surviving analysis from antiquity of the complex of issues raised by the notion of chrematistics (and no reason to think that other systematic accounts once existed and were eventually lost). That is the account by Aristotle and it is not only unsatisfactory but there can be no doubt that Aristotle himself thought it merely tentative and incomplete. It is enough to indicate that the term 'chrematistics' is used in three incompätible senses, with the pejorative one predominant (see in detail Newman 1887–1902, II, 165–208). Nevertheless, Aristotle had made a serious start in grappling with a major social and moral problem, and one could have expected further discussion and development. Instead, the whole discussion promptly died for some 1,500 years until it was apparently revived with Albertus Magnus and Thomas Aquinas in the thirteenth century following the translation into Latin of the *Ethics* and the *Politics.* This is a curious puzzle in the history of ideas, and it must be said that modern accounts are dominated by illusions.

One is that not only was Aristotle one of the greatest figures in the history of philosophy, but that this was acknowledged in the generations after his death by the way in which philosophical questions were treated. Perhaps this was not often stated explicitly (since there was no textual foundation for such a view), but the implication was never far beneath the surface. Yet it is a fact that in the field of social thought the *Politics* was effectively lost sight of and unknown for some three centuries after the death of Aristotle's successor Theophrastus (see Sandbach 1985), and that it continued to be neglected for a millennium. Between Cicero and the thirteenth century no more than a dozen references in Greek are known, half of them in Byzantine lexica and scholia (Susemihl and Hicks 1894, p. 18 n.7). Stoicism had quickly become the dominant school and the concept of natural law was its pivotal one. From the Greeks it passed to the Romans, and under the influence notably of Cicero and Seneca, it became one of the three foundation-stones of European ethics until the beginning of the modern era (and even beyond). How 'pure' this Stoic ethics remains is irrelevant. What is essential is its dominance and the fact that it by-passed the Aristotelian concern with chrematistics, justice in exchange, and the like.

The second foundation-stone was the Roman law, a confusing and ambiguous one. The Roman jurists were in practice firm supporters of the principle that any agreements made in good faith and not specifically prohibited by the law were acceptable. There was no place in that sphere for Aristotelian doctrines about chrematistics. Nor was there in the third foundation, the Church Fathers, despite their 'misgivings about the merchant' (Baldwin 1959, pp. 12–16). The attitude of the New Testament to economic matters, unlike the Old, was a subject of controversy from the beginning, but what matters in our context is first that there was a hardening of attitudes on some matters, especially usury and the morals of commerce, over the final centuries of antiquity and the early Middle Ages, and that secondly, chrematistics hardly entered the discussion. From the sixth century on, church councils and later, in the Carolingian period, imperial enactments produced a stream of warnings about the 'morally dangerous character of buying and selling' (Baldwin 1959, p. 34). The fullest account of this material remains Schaub (1905), with its revealing subtitle, 'eine moralhistorische Untersuchung'. By the twelfth century, an impressive body of doctrine had been developed regarding usury, business practices and trade. The contrast is striking with the almost total emptiness on these subjects in the vast corpus of the writings of St Augustine (Cranz 1954).

It is remarkable how the conventional histories of economic thought tend to skip over some seven centuries of 'moral history' as they leap from antiquity to the rediscovery of Aristotle and the apogee of scholasticism in the thirteenth century. In an important and neglected article published by the Pontifical Medieval Institute in Toronto, Eschmann (1943, p. 134) rightly concluded that in that century, dominated by Albertus Magnus and Thomas Aquinas, the rediscovery of Aristotle

merely 'reinforces the Roman–patristic ideas and plays a decorative rather than a constructive role'. By then the fundamental principles had already been laid down in the solution of concrete problems arising from usury, buying and selling, and the conduct of trade. Aquinas may have codified such doctrines as just price, but he neither invented them nor found them in Aristotle. Because of the way he worked, in particular because he never produced a synthesis of social philosophy, there are puzzles that continue to plague commentators.

How, in particular, was the just price determined? That was Aristotle's problem (though in different language) with chrematistics, and, as we have seen, he never resolved it satisfactorily. But at least he made a serious effort: we know of none made by Aquinas. The common view today is that for the latter the just price was the prevailing one in normal practice, and I cannot avoid the suspicion that this is merely a projection into scholasticism of the thinking of neoclassical economists (e.g., Viner 1978). As one historian who accepts the common view has conceded, nowhere did Aquinas 'state in practical terms what exactly comprised the just price' and the only textual support for the notion that the just price is the prevailing price comes from writers in the next generation or two (Baldwin 1959, pp. 75–6). There is an alternative interpretation that at least merits more consideration than it has hitherto received: Aquinas, writes a neo-Thomist economist (Stark 1956, p. 5), 'does not stop to consider the question how the just price is arrived at' because 'it is taken for granted . . . like all the other rules and regulations of an orderly social existence'; 'the price is part and parcel of the system of custom on which all social life is built'.

Ironically, at the moment when the Aristotelian questions in their latter-day versions reached their intellectually most sophisticated formulation they were already moribund in practice, and they soon effectively dropped from sight, despite the lingering canon law and theological interest, in usury in particular. The slogan 'treasure by foreign trade' can be thought to have been the death-knell.

See Also

▶ Aristotle (384–322 BC)

References

Baldwin, J.W. 1959. The medieval theories of the just price. *Transactions of the American Philosophical Society* 49(4): 1–92.

Cranz, F.E. 1954. The development of august ideas on society. *Harvard Theological Review* 47(4): 255–316.

Eschmann, I.T. 1943. A thomistic glossary on the principle of the preeminence of the common good. *Mediaeval Studies* 5: 123–165.

Finley, M.I. 1970. Aristotle and economic analysis. *Past and Present* 47: 5–25.

Newman, W.L. 1887–1902. *The politics of Aristotle,* 4 vols. Oxford: Oxford University Press.

Sandbach, F.H. 1985. *Aristotle and the stoics*, Supplement, vol. 10. Cambridge: Cambridge Philological Society.

Schaub, F. 1905. *Der Kampf gegen den Zinswucher, ungerechten Preis und unlautern Handel in Mittelalter.* Freiburg: Herder.

Stark, W. 1956. The contained economy. Paper no. 26 of the Aquinas Society of London.

Susemihl, F., and R.D. Hicks (eds.). 1894. *The politics of Aristotle, Books I–V.* London: Macmillan.

Viner, J. 1978. Religious thought and economic society. *History of Political Economy* 10(1): 9–189.

Christaller, Walter (1894–1975)

Gordon C. Cameron

Keywords

Central place theory; Christaller, W.; Market threshold; Normal travelling distance

JEL Classifications

B31

Christaller, who never held an academic post but worked throughout his life in association with the University of Erlangen, is known for one seminal book *Die zentralen Orte in Süddeutschland* [Central Places in Southern Germany]. Published in Germany in 1933 it remained largely unnoticed

by English-speaking scholars until a translation of August Lösch's *Economics of Location* (1954) brought it widespread attention. Later an accurate translation of Christaller's book by C.W. Baskin (in 1966) confirmed the elegance of his deductive theorizing.

Christaller sought to clarify and explain the laws which determine the number, sizes and distribution of towns. Drawing upon the work of von Thünen, Alfred Weber and Engländer, Christaller developed a general theory of why a *hierarchy* of villages and towns providing different services should appear and why this hierarchy should differ region by region. Making use of key concepts of market threshold, and normal travelling distance, he showed how the geographical extent of the trading areas for different goods and services vary and how low order centres provide limited ranges of goods to small trading areas whereas larger centres service much wider areas and contain all the goods of the lower centres as well as goods unique to their size.

Christaller's work has been criticized as ignoring the role of manufacturing in shaping the growth of towns and cities, of underplaying the effects of an unequal distribution of natural resources and of an all too rigid expression of the laws of market size and of the hierarchy of central places. Of the last point Christaller was fully aware and by 1950 he had modified his stance allowing for greater variability in the determinants of the hierarchy. And though his general theory of spatial relations is incomplete, all subsequent analysts of retail trade, of the location of services and of urban growth, recognize the rigour of his approach and the elegance of his attempt to provide the 'economic theoretical foundations of town geography'.

See Also

► Central Place Theory

Selected Work

1933. *Die zentralen Orte in Süddeutschland: eine ökonomisch-geographische Untersuchung uber die Gesetzmässigkeit der Verbreitung und Entwicklung der Siedlungen mit stadtischen Funktionen*. Jena: G. Fischer Trans. C.W. Baskin as Central Places in Southern Germany, Englewood Cliffs: Prentice-Hall, 1966.

Bibliography

Berry, B.J.L., and C.D. Harris. 1970. Walter Christaller: An appreciation. *Geographical Review* 60: 116–119.

Lösch, A. 1954. *The economics of location.* Trans. from the 2nd ed., 1944 by W.H. Woglom with the assistance of W.F. Stolper, New Haven: Yale University Press. 1st German ed., 1940.

Christian Socialism

E. Cannan

Christian Socialism is a name which properly belongs to the propagation of cooperative production or working men's associations by F.D. Maurice and his disciples in the years 1849 to 1853. Its origin is to be found in a letter from J.M. Ludlow to Maurice (March 1848) saying that the socialism of Paris workmen was a real power which would shake Christianity if it were not Christianized. After the publication of Henry Mayhew's letters on the London poor in the *Morning Chronicle,* in 1849, Maurice and his followers at Lincoln's Inn, who had already been trying to persuade the Chartists, in *Politics for the People* (6 May to 29 July 1848), and in discussions at the Cranbourne Tavern, that moral and sanitary reform were of much more importance than extension of the suffrage, turned their attention to economic questions. They were led to deny any beneficence to the operation of self-interest. 'Free competition', said Ludlow, 'mars every-where, instead of making, the wisest distribution of labour' (*Christian Socialism,* p. 35). 'We have protested', Maurice wrote to Dr Jelf, 12 November 1851, 'against the spirit of competition and rivalry precisely because we believe it

is leading to anarchy, and must destroy at last the property of the rich as well as the existence of the poor' (*Life*, ch. ii, p. 83). As a remedy they proposed 'Christian socialism', or friendly association for productive purposes. They sometimes went so far as to imagine a state of things in which all producers might 'combine regularly into one body which should, after mutual explanations and by mutual concert, fix the terms upon which each member should dispose of his wares to the others' (Ludlow, *Christian Socialism,* p. 35); but they suggested no principle of distribution on which this agreement should be based. They founded an association of tailors (February 1850) of which Walter Cooper, formerly a Chartist, was manager, and organized a society for promoting working men's associations under a council of promoters among whom were Maurice, Charles Kingsley, T. Hughes, E.V. Neale, and F.J. Furnivall. *Alton Locke,* which represents the ethical side of the Christian Socialist doctrine, was published early in 1850, and was followed by *Tracts on Christian Socialism, Tracts by Christian Socialists*, and the *Christian Socialist*, a weekly penny paper which lasted from 20 November 1850 to the end of 1851. Its place was then taken by the *Journal of Association,* which endured till 28 June 1852. The evidence of the 'Promoters' before Slaney's Committee of the House of Commons on 'Investments for the savings of the middle and working classes' in 1850, aided in bringing about the legislation of cooperative societies by the 'Industrial and Provident Partnerships Act' of 1852. After the passing of that Act the society for promoting associations was remodelled and the term 'Christian Socialism', as employed in this connection, was abandoned. It was offensive alike to theologians, economists, and socialists. The hostility displayed towards the Christian Socialists in many quarters was more due to the name they assumed, and to the vehemence with which Kingsley denounced competition, than to dislike of their Associations, though these were doubtless looked on with some suspicion as copies from French models.

References

Brentano, L. 1883. *Die christlichsoziale Bewegung in England*, 2nd edn. Leipzig, 1883.

Hughes, T. 1884. *Prefatory Memoir in the Eversley edn of Alton Locke,* 2 vols. London: John Murray, 1881.

Ludlow, J.M. 1851. *Christian socialism and its opponents*. London.

Maurice, F. 1884. *Life of Frederick Denison Maurice, chiefly told in his own letters*. London.

Circular Flow

Giorgio Gilibert

JEL Classifications
E1

The analysis of the social process of production and consumption must start from some notion of commodity circulation. Consideration of the simple cycle of agricultural production suggests that production is an essentially circular process, in the sense that the same goods appear both among the products and among the means of production. From this viewpoint, commodity (as well as money) circulation is a triviality, whose discovery cannot really be attributed to any particular economist.

It has been suggested that the notion was originally developed by François Quesnay, a surgeon, by analogy with the circulation of the blood. However the popular analogy between money and blood is much older (see for instance 'Money is for the state what blood is for the human body', *Etats généraux,* 1484); and the process of money and commodity circulation among different classes (landlords, labourers, merchants) and areas (town and country) was clearly described by Boisguillebert and Cantillon several decades before the physiocrats.

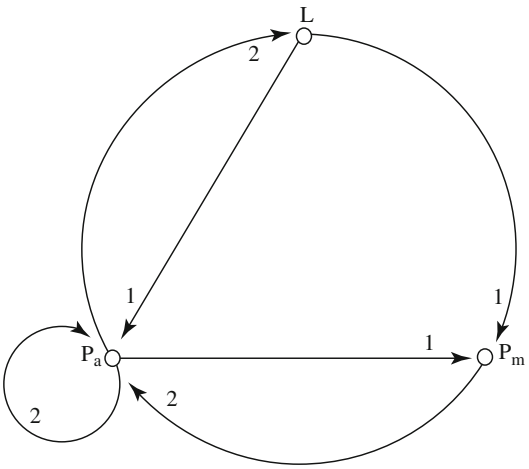
What is truly novel with Quesnay is the idea that the essential task of economic science is the

investigation of the technical and social conditions which allow the repetition of the circular process of production. This approach (at least in the extreme form given it by the physiocrats), and the peculiar model building activity that sprang from it, was later abandoned by economists. More than a century had to pass before the theme could be resumed, following the publication of Marx's own *tableaux* in the second volume of *Capital* (Marx 1885), but merely within the rather limited and isolated group of the German and Russian theoretical economists.

Tugan-Baranowsky considered circularity as the essential feature of capitalist economy, in which production was the end of consumption rather than the other way round; in his view, the economists were unable to understand this 'paradox' because (with the remarkable exception of Marx) they had strayed from the way opened up by Quesnay. The young Schumpeter, in a justly celebrated essay, dated the birth of economics as a science from the physiocratic analysis of the circular flow. And Leontief (1928) wrote in a similar vein, arguing in favour of the substitution of the principle of circular flow (the 'reproducibility viewpoint') for that of *homo oeconomicus* (the 'scarcity viewpoint') as the cornerstone of economic theory.

The reproducibility viewpoint is shared by the whole classical tradition of political economy. However, within this broad theoretical tradition, we can single out a radical strand which considers the economic behaviour of every individual as completely determined by the reproduction requirements of the system. This peculiar approach characterizes the pure theorists of the circular flow, with whom we will now briefly deal. Not surprisingly, this theoretical approach is often associated with a practical attitude in favour of some sort of central planning (as a consequence of the distrust for the 'anarchy' of the market).

The *Tableau Economique* depicts all the transactions taking place during the year among the three basic classes of society: the class of landowners (L), the 'productive' class of farmers (P_a), and the 'sterile' class of manufacturers (P_m). These transactions can be summarized by a graph, where three points – one for each class – are connected by lines, representing the transactions; the lines are oriented according to the direction of the money flows, whose value is shown by numbers (thousand millions of *livres*). Figure 1 is drawn on the data of Quesnay (1766); since the sum of the money flows leaving each point equals that of those coming in, the system is reproducible.

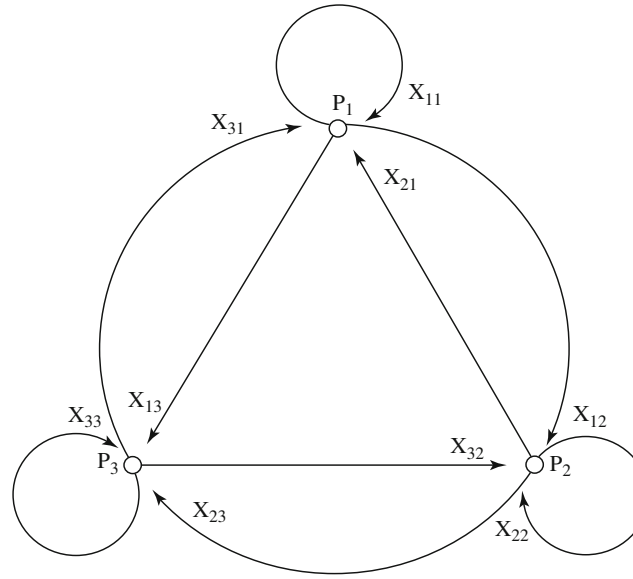


Circular Flow, Fig. 1

Marx's (simple) reproduction scheme can also be easily adapted to the same type of three-point graph, once capitalists are substituted for landowners, and the two industries producing intermediate goods ('constant' capital) and consumption goods ('variable' capital and luxuries) are substituted for the two classes of manufacturers and farmers respectively. It should be noted that, while Quesnay's *tableaux* are inherently static, Marx does also consider expanded reproduction: in his own words, the picture shifts from a circle to a spiral. A modern example of a circular representation of an expanding economy is the well-known von Neumann model, which, from this point of view, can be considered as the most sophisticated heir to the Marxian schemes.

Quesnay's and Marx's *tableaux* were offered in value terms; but there is no conceptual difficulty in imagining analogous schemes in physical terms. Now, if all the physical transactions taking place

Circular Flow, Fig. 2



among all the agents of the economy are known, there is a unique set of relative prices which makes it possible for the process to be repeated.

Let us consider an economy in which n producers produce n goods. If we know all the physical amounts x_{ij} of the various goods consumed by the different producers, and if the economy is closed (i.e. production equals consumption for each good), relative prices p_i are determined by the following linear homogeneous equations:

$$\sum_i x_{ij} p_i = p_j \sum_h x_{jh}.$$

This theory or prices has now come to be associated with the closed Leontief model (Leontief 1941), but it was originally formulated in the late 18th century by Achille Isnard. He considered a simple example with three producers and consistently computed the corresponding prices.

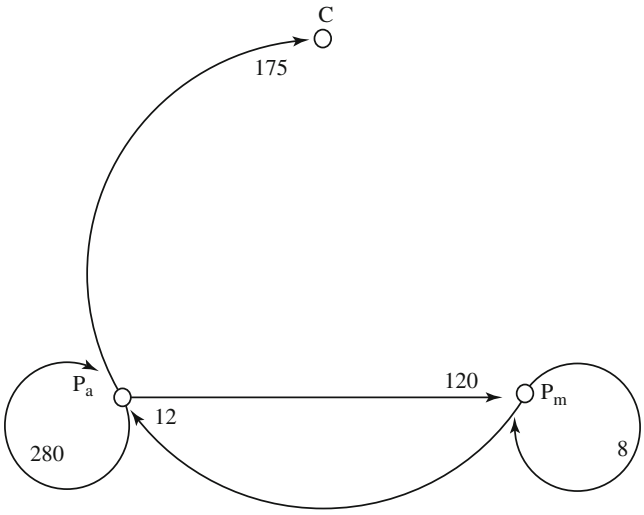
His example is illustrated by the graph of Fig. 2: three points, one for each producer, are connected by lines, corresponding to the physical amounts exchanged; the lines are now oriented according to the physical commodity flows. Relative prices have to be such as to equalize the value of the flows leaving each point with that of the flows coming in; the loops at the vertices (self-consumption) are not relevant to our problem.

When Leontief, a century and a half later, rediscovered the theory, he recognized in it the 'objective' theory of value. One year later, the German mathematician Robert Remak interpreted system (1) as determining the rational prices for an economy in which the individual standards of living are fixed by a central authority. He showed that the system has in general meaningful solutions; and maintained that these prices could be practically computed and implemented.

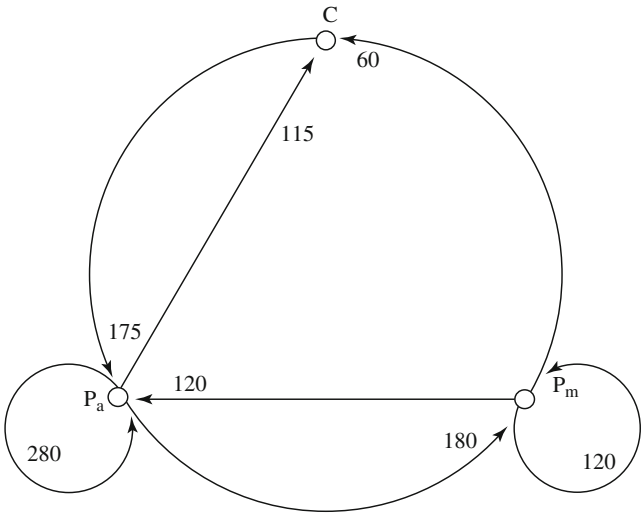
Until now, we have considered only closed systems, in which all transactions are assumed as known irrespective of their nature (technical inputs or human 'final' uses). We can now open the model, by considering as given only those transactions which are dictated by the technology in use (including workers' subsistence) and leaving undetermined the final utilization of the surplus thus appearing.

There is now room for an additional relation, stating the way in which the surplus is distributed. If we assume that it is entirely appropriated by profit-earners in proportion to the capital advanced, we land on the familiar ground of the classical theory of production prices.

Circular Flow, Fig. 3



Circular Flow, Fig. 4



The case can be illustrated by a simple numerical example supplied by Sraffa: there are only two industries, producing wheat (P_a) and iron (P_m) respectively; the class of capitalists (C) gets the entire surplus, consisting only of wheat. In Fig. 3 the numbers on the oriented graph refer to the physical quantities (quarters and tons) in the example.

The uniform profit rate has to be such as to equalize the value of the surplus bought by capitalists to the profits accruing to them; and the exchange value between the two commodities has to be such as to enable each industry to replace its advances and to distribute profits in proportion to their value. Loops are now relevant.

The system is then reproducible when the money flows leaving each point are equal to those coming in; the situation is illustrated in Fig. 4, and corresponds to a price of iron in terms of wheat equal to 15 and to a common profit rate equal to 25 per cent.

Finally, if we allow the wage earners to share the surplus with the capitalists, we generate the pure theory developed by Piero Sraffa (1960).

We are now able to interpret the abstract transition from our original circular theory to the classical theory of production prices, and eventually to its modern Sraffa version, as successive steps in a gradual opening of the model. From an initial system in which the economic behaviour of every individual is assumed to be rigidly determined by reproduction requirements, we have passed to a system in which capitalists (and rentiers) are assumed to be free in determining their final demand; and finally we have also granted some degree of freedom to the workers.

The term 'free' means here only that the composition of final demand is an issue which lies outside the domain of the pure theory of prices; of course, it can be the object of a distinct section of economic theory. In this perspective, we could say that the neoclassical theory of prices corresponds to a vision of the economy in which the individuals are supposed to be undifferentiated (i.e. there are no classes) and all equally free (the reproduction requirements do not play any essential role in determining prices).

See Also

- ▶ Physiocracy
- ▶ Quesnay, François (1694–1774)

Bibliography

Cantillon, R. 1755. *Essai sur la nature du commerce en général*, 1952. Paris: INED.

de Boisguillebert, P.. 1707. Dissertation de la nature des richesses. In *Oeuvres manuscrites et imprimées.* Paris: INED, 1966.

Isnard, A.N. 1781. *Traité des richesses*. Lausanne: Grasset.

Leontief, W. 1928. Die Wirtschaft als Kreislauf. *Archiv für Sozialwissenschaft und Sozialpolitik* 60 (3): 577.

Leontief, W. 1941. *The structure of American economy.* Oxford: Oxford University Press.

Marx, K. 1885. *Das Kapital*. Vol. II. Hamburg: Meissner.

Neumann, J. von. 1937. Über ein ökonomisches Gleichungssystem und eine Verallgemeinerung des Brouwerschen Fixpunktsatzes. Ergebnisse eines mathematischen Kolloquiums VIII.

Peter, H. 1954. *Mathematische Strukturlehre des Wirtschafts-kreislaufes*. Göttingen: Schwartz.

Quesnay, F. 1766. Analyse de la formule arithmétique du Tableau économique. In *Textes annotés*. Paris: INED, 1958.

Remak, R. 1929. Kann die Volkswirtschaftslehre eine exakte Wissenschaft werden? *Jahrbücher für Nationalökonomie und Statistik* 76.

Remak, R. 1933. Können superponierte Preissysteme praktisch berechnet werden? *Jahrbücher fü Nationalökonomie und Statistik* 80: 839.

Schumpeter, J. 1914. Epochen der Dogmen und Methodengeschichte. In *Grundriss der Sozialökonomik, Tübingen: Mohr.* Trans. R. Aris as Economic Doctrine and Method: An historical sketch. Oxford: Philip Allen, 1954.

Sraffa, P. 1960. *Production of commodities by means of commodities*. Cambridge: Cambridge University Press.

Tugan-Baranowsky, M. 1894. Les crises industrielles en Angleterre. French trans., Paris: Giard, 1913.

Circulating Capital

Mark Blaug

Abstract

This article summarizes the history of the distinction between circulating capital (whose full value returns to the capitalist from the sale of final goods) and fixed capital (whose value is never fully recovered in one production cycle) from its introduction by Smith and development by Ricardo to its treatment by Marx and the Austrian capital theorists. It gave rise to the wages fund doctrine, the problem of joint production, and the issue of the optimum rate of depreciation and replacement of old equipment.

Keywords

Advances; Austrian capital theory; Böhm-Bawerk, E. von; Capital theory; Circulating capital; Constant and variable capital; Depreciation; Fixed capital; Frisch, R. A. K.; Joint production; Marginal revolution; Marx, K. H.; Mill, J. S.; Quesnay, F.; Ricardo, D.; Smith, A.; Turgot, A. R. J.; Wages fund; Wicksell, J. G. K.

JEL Classifications
E22

The explicit distinction between fixed and circulating capital first makes its appearance in Book II, chapter 1 of Adam Smith's *Wealth of Nations,* who derived it from ample hints in Quesnay and Turgot. Circulating capital goods, according to Smith, consist of those intermediate goods that embody a quantity of purchasing power that perpetually returns to the capitalist as he disposes of the final goods into the making of which they entered, in contrast to fixed capital goods, whose value is never fully recovered in one production cycle. The simplest example of circulating capital is raw materials, just as the simplest example of fixed capital is buildings and machines. However, all the classical economists, including Smith, included in circulating capital not just raw materials but also the consumer goods that support labour during the process of production; that is, wage goods.

This is the origin of the notorious 'wages fund doctrine', according to which wages are said to be 'advanced' to workers at the outset of a production period as a result of which they are determined by the ratio between the volume of capital advanced and the size of the labour force. The notion arose out of a pronounced tendency in 18th-century economics to regard agriculture as an industry typical of production as a whole and to view wheat as both a representative output of agriculture and the staple article of consumption of workers. The fact that wheat only becomes available in the form of annual harvests, which must be willy-nilly stored as a 'fund' for future consumption if its actual use is to be more or less continuous throughout the year, made it possible to define capital simply as 'advances' to workers to support them from seed-time to harvest. Despite the fact that this agrarian model was gradually abandoned in the century after Smith, the wages fund doctrine lived on until J.S. Mill's recantation of the doctrine in 1867, and with it the definition of circulating capital as including all consumer goods that enter into the wage basket (Blaug, 1985, pp. 185–8). Surprisingly enough, this conception of capital as consisting largely if not solely of wage goods survived even beyond the 'marginal revolution': it lies at the heart of the theoretical schema adopted by Böhm-Bawerk in his *Positive Theory of Capital* (1887).

Adam Smith noted that fixed and circulating capital combine in different proportions in different industries, but it was Ricardo who converted this observation into one of the central facts of industrial life in a capitalist economy and a major problem for the theory of value. Ricardo wanted to argue that relative prices are determined by relative labour costs but, as he candidly admitted in the first chapter of the *Principles of Political Economy and Taxation*, this cannot be true, because not only does the ratio of fixed to circulating capital differ between industries but, in addition, the two kinds of capital may differ in durability between industries. Indeed, he added in a footnote, the distinction between fixed and circulating capital is not essential because any difference between them is solely a matter of degrees of durability; that is, the different time periods for which capital is locked up in the productive process: circulating capital is the sum of goods tied up in production for only as long as the period of production in question, whatever its length, whereas fixed capital is a joint output of this production period in the shape of a slightly older building or a slightly older machine. To put it in a nutshell: the distinction between fixed and circulating capital is not the difference in their absolute durability but rather the difference in their durability relative to the length of the production period in which they are employed.

Thus, despite the fact that Marx in *Capital* rejected the Smithian distinction between fixed and circulating capital and chose instead to distinguish 'constant' and 'variable' capital, confining the former to the wage bill and the latter to everything else on the grounds that wages might vary for a given production system even if all the technical input coefficients remained the same, he operated throughout the first volume of the book with a circulating capital model by virtue of the assumption that the capital stock of every industry in the economy turns over once a year: despite all the references to machinery in this first

volume, all the analytical problems created by the use of fixed capital are eliminated by assuming that every industry operates with an annual production period. It is only in volume 2 of *Capital,* and particularly chapters 8–14, that Marx takes account of differences in the durability or turnover rates of capital invested in different industries, and it is here that he begins to confront the problems created by the fact that fixed capital, unlike circulating capital, only transfers part of its value to the final product during each turnover of capital. This is the now famous problem of joint production, which, it has been argued (Steedman 1977, ch. 10), may produce such anomalies as negative labour-costs for some products.

In the same way, all of the work of Böhm-Bawerk and most of that of Wicksell on the theory of capital is confined to the question of the optimum investment period of continuously applied circulating capital; that is, to what Ragnar Frisch has called the 'flow input–point output' case. It is only when we take up the 'point input–flow output' or the even more typical case of 'flow input–flow output' that we confront the question of fixed capital, an issue that Böhm-Bawerk consistently avoided and that Wicksell only took up in one essay in later life (Blaug 1985, pp. 563–4). The difficulty created by the use of fixed capital is simply that there is no obvious way of linking particular units of input embodied in fixed capital with particular units of finished output: all the inputs embodied in fixed equipment are jointly responsible for the whole stream of future outputs. Thus, by limiting itself to circulating capital, Austrian capital theory avoided such vexing questions as the optimum rate of depreciation and replacement of old equipment that are always linked with the decision to invest in new equipment, questions which perhaps are not completely resolved even to this day.

The increasing use of fixed capital is said to be one of the distinguishing characteristics of a capitalist system. If so, we might well expect capital theory to have been largely devoted to an analysis of fixed capital. It is one of the ironies of the history of economic thought, however, that capital theory from Turgot to the late Wicksell always treated circulating and not fixed capital as 'capital' *par excellence.*

Bibliography

Blaug, M. 1985. *Economic theory in retrospect*. 4th ed. Cambridge: Cambridge University Press.

Steedman, I. 1977. *Marx after Sraffa*. London: New Left Books.

City and Economic Development

J. Vernon Henderson

Abstract

As countries develop they urbanize, with resources shifting from labour-intensive agricultural production to manufacturing and services, which are located in cities because of agglomeration economies. This entry discusses the economic determinants of this process. But urbanization also moves populations from traditional rural environments with informal political and economic institutions to the relative anonymity and more formal institutions of urban settings. A major issue in the development process is development of institutions and national policies which allow cities to operate in markets that are well structured and conducive to good urban outcomes.

Keywords

City in economic development; Congestion; Core-periphery models; Democracy; Endogenous growth; Human capital; Korea; Land tax; Regional models; Rent seeking; Rural-urban migration; Spatial convergence; Two-sector models; Urban agglomeration; Urbanization

JEL Classification

O18

The city in economic development is fundamental to the urbanization process. Urbanization, or the shift of population from rural to urban environments, is a transitory process which is socially and

culturally traumatic. As a country develops, it moves from labour-intensive agricultural production to labour being increasingly employed in industry and services. The latter are located in cities because of agglomeration economies. Thus, urbanization moves populations from traditional rural environments with informal political and economic institutions to the relative anonymity and more formal institutions of urban settings. That in itself requires institutional development within a country.

Once urbanization is complete, one might be tempted to simply move on to the traditional analysis of systems of cities, with the idea that the issues that face systems of cities in developed economies are the same as those that face cities in developing but fully urbanized economies (as in Latin America and the Middle East). But in practice this is not the case; countries still face problems of developing institutions and national policies which allow cities to operate in markets that are well structured and conducive to good urban outcomes. Here, we discuss both the urbanization process and then the institutional-policy issues that face cities in developing countries.

The Urbanization Process

There are several models of the urbanization process. The traditional ones are two- sector models, where population moves from a rural sector to an all-purpose urban sector, due to exogenous factors such as unexplained shifts in technology (Lewis 1954). Dual-sector models focus on the question of urban 'bias', or the effect of government policies on the urban–rural divide, and the efficient rural–urban allocation of population at a point in time. Generally, these models are static, and any urbanization is the result of exogenous forces – technological change favouring the urban sector or changes in the terms of trade favouring the urban sector. There is a new generation of two-sector models, namely, the core–periphery models, which have more of a spatial flavour (Krugman 1991; Puga 1999). Core–periphery models ask when in a two-region country industrialization, or 'urbanization', is spread over both regions rather than being concentrated in just one region. The models explore a key issue: the initial development of a core (say, coastal) region and a periphery (say, hinterland) region, as technology improves (transport costs fall) from a starting point with two identical regions. However core–periphery models have limited implications for urbanization per se. They are unidimensional in focus, asking what happens to core–periphery development as transport costs between regions decline; they are really regional models, with limited urban implications. Urban models are focused on the city formation process, where the urban sector is composed of numerous cities, endogenous in number and size. Efficient urbanization and growth require timely formation of cities and appropriate institutions.

Henderson and Wang (2005) develop an endogenous growth model with accumulation of human capital, where there is a shift out of the rural sector into an urban sector as per capita human capital and income grow. The urban sector is composed of multiple cities which grow in size with knowledge accumulation and in numbers with national population growth and rural–urban migration. Urbanization occurs because demand for food products is postulated to be income inelastic, so as per capita incomes rise the relative demand for food products declines, while at the same time productivity in the rural sector is growing. That releases labour from the rural sector to migrate to the urban sector, where the relative national demand for urban products is rising overtime.

As the urban sector grows, new cities form in national land markets. Efficient city sizes are limited, reflecting a trade-off between marginal agglomeration economies as a city grows and steadily rising urban diseconomies in the form of commuting, congestion and other urban disamenities. Efficient city sizes are at or near the peak to each city's inverted-U shape relationship between real income per worker and city employment where, with economic growth, such peaks and efficient city sizes may be shifting out over time. With urbanization and national population growth, if existing cities are to stay near efficient sizes, new cities need to form in a timely fashion. That timely formation requires local

governments to have the autonomy to tax land rents and exclude entrants through zoning provisions. Moreover, developers or local governments must have the autonomy to utilize land and undertake enormous urban infrastructure investments so as to form new large- scale settlements. Such institutions and market environments may not be in place or may be slow to develop, and national politics may delay their evolution, especially in developing countries. These factors retard the timely formation of cities, forcing migrants into existing oversized cities. We discuss these issues below.

Empirics and Policy Issues

The policy and empirical literature on urbanization addresses three broad questions. These deal with the determinants of the rural–urban allocation of resources at any point in time, spatial convergence, and excessive urban concentration.

Rural–Urban Allocation of Resources

Dual economy models in the traditional development literature ask whether market failures bias the allocation of resources between the urban and rural sectors or between bigger and smaller cities. Renaud (1981) makes the related point that it is not just market failures but explicit government policies that bias or influence urbanization through their effect on national sector composition. Policies affecting the terms of trade between agriculture and modern industry or between traditional small town industries (textiles, food processing) and high-tech large city industries affect the rural–urban or small–big city allocation of population. Such policies include import tariffs, price controls and product subsidies.

Spatial Convergence

The issue of convergence across spatial units in a country was initially posed at the regional level. Williamson (1965) argued that national economic development is characterized by an initial phase of internal regional divergence of per capita incomes and the allocation of industrial resources, followed by a phase of later convergence. There is a related urban model of this divergence–convergence phenomenon, which looks at urban primacy and the quantity allocation of resources across cities. Following Ades and Glaeser (1995), conceptually the urban world is collapsed into two regions: the primate city versus the rest of the country, or at least the urban portion thereof. The question is: to what extent is urbanization concentrated in, or confined to, one (or a few) major metro areas, as opposed to being spread more evenly across a variety of cities? Primacy is commonly measured by the ratio of the population of the largest metro area to the entire urban population in the country. Ades and Glaeser (1995) and Davis and Henderson (2003) find that primacy first increases, peaks, and then declines with economic development, indicating a later spread of urban resources from the primate city to other cities over time.

As part of this spatial convergence process, Lee (1997) and Kolko (1999) explore the relationship between changes in urban concentration and industrial transformation for Korea since 1975 and for the USA since 1900. The idea is that manufacturing is first concentrated in primate cities at early stages of development, and then decentralizes to such an extent that at the other end of economic development it is relatively more concentrated in rural areas. Initial concentration fosters ‘incubation’ and adaptation of technologies from abroad in a concentrated urban environment. But once manufacturing has modernized with fairly standardized technologies, firms decentralize to hinterland locations where rent and wage costs are cheaper. For example, in Korea Seoul’s urban primacy peaked around 1970, when Seoul had a dominant share of national manufacturing. During the next 10 or 15 years, manufacturing suburbanized from Seoul to nearby satellite cities, as well as to satellite cities surrounding the two other major metro areas, Pusan and Taegu. But then in the early 1980s manufacturing spread rapidly from the three major metro areas and their satellites to rural areas and other cities. The largest metro areas became business service-intensive, relying on economies of diversity in local business services, often purchased by headquarter units of

firms as part of marketing, financing, and exporting activities for their goods produced by plants in hinterland locations. This spatial separation, with headquarters' activities of firms in large metro areas and production facilities in smaller specialized cities, is called 'functional specialization' by Duranton and Puga (2005).

Urban Concentration

A third set of questions asks whether the degree of urban concentration in countries is too little or too much. Are there policies which bias development towards bigger, say, politically dominant coastal cities at the expense of smaller, say, hinterland cities? The basic idea is that the political system favours the national capital (or other seat of political elites such as São Paulo in Brazil). For example, direct restraints on trade for hinterland cities such as an inability to access capital markets or to get export or import licences favour firms in the national capital. Policymakers and bureaucrats may gain as shareholders in such firms, or they may gain rents from those seeking licences or other exemptions from trade restraints. Indirect trade protection for the primate city can also involve underinvestment in hinterland transport and communications infrastructure. Another strategy can be to retard development of institutions and national land markets that allow timely formation of large-scale, competitor hinterland cities. Whether as true beliefs or as a cover for rent-seeking behaviour, policymakers often articulate the view that large, favoured cities are more productive and thus should be the site for government-owned heavy industry (such as São Paulo or Beijing–Tianjin, historically). Unfortunately these heavy industries don't benefit sufficiently from the agglomeration economies in such large cities and can't afford their higher costs of land and labour, which is one reason why they lose money in such cities.

Favouritism of a primate city creates a non-level playing field in competition across cities. The favoured city draws in migrants and firms from hinterland areas, creating an extremely congested high-cost-of-living metro area. Local city planners can try to resist the migration response to primate city favouritism by, for example, refusing to provide legal housing development for immigrants or to provide basic public services in immigrant neighbourhoods. Hence squatter settlements, bustees, kampongs and so on may develop. But still, favoured cities tend to draw in enormous populations.

What is the econometric evidence indicating that politics plays a role in increasing sizes of primate cities? Based on cross-section analyses, Ades and Glaeser (1995) find that, if the primate city in a country is the national capital, it is 45% larger. If the country is a dictatorship, or at the extreme of non-democracy, the primate city is 40–45% larger. The idea is that representative democracy gives a political voice to hinterland regions, so limiting the ability of the capital city to favour itself; and fiscal decentralization helps level the playing field across cities, giving hinterland cities political autonomy to compete with the primate city. Davis and Henderson (2003) explore these ideas further, examining in a panel context the impact of democratization and fiscal decentralization upon primacy. Examining democratization and fiscal decentralization together, they find moving from most to least democratic form of government reduces primacy by 8%, and moving from most to least centralized government reduces primacy by 5%. They also find transport infrastructure investment in hinterlands reduces primacy, a prediction of core–periphery models.

Given the urban primacy relationships, it is natural to ask whether urban concentration is important to growth. Is there an optimal degree of urban primacy with each level of development where significant deviations from this level detract from growth? Optimal primacy would involve a trade-off between the benefits of increasing primacy (enhanced local scale economies contributing to productivity growth) and the costs (more resources diverted away from productive and innovative activities to shoring up the quality of life in congested primate cities). Henderson (2003) examines this question with panel data methods and finds that there is an optimal degree of primacy at each level of development which maximizes national productivity growth. That optimal degree rises as country income declines: high relative agglomeration is important

when countries have low knowledge accumulation, are importing technology, and have limited capital to invest in widespread hinterland development. There is an international tendency to excessive primacy, with effectively non-federated countries such as Argentina, Chile, Peru, Thailand, and Algeria having extremely high primacy.

While for countries where people are allowed to migrate freely across cities and from rural to urban areas the focus is on excessive urban concentration, in the former planned economy countries the concern goes the other way. Countries such as China have formal migration restrictions limiting the visas given to rural people to move to cities and limiting migrants' access to jobs, housing, medical care and schooling in destination cities to reduce the incentive to migrate. Other former planned economies primarily limited migration through restrictions on housing provision and land development in cities. Planned economies have much lower urban concentration than other large countries. The efficiency loss there derives from unexploited urban agglomeration economies.

See Also

- Location Theory
- Spatial Economics
- Systems of Cities
- Urban Agglomeration
- Urbanization
- Urban Production Externalities

Bibliography

Ades, A., and E. Glaeser. 1995. Trade and circuses: Explaining urban giants. *Quarterly Journal of Economics* 110: 195–227.

Davis, J., and J.V. Henderson. 2003. Evidence on the political economy of the urbanization process. *Journal of Urban Economics* 53: 98–125.

Duranton, G., and D. Puga. 2005. From sectoral to functional urban specialization. *Journal of Urban Economics* 57: 343–370.

Henderson, J. 2003. The urbanization process and economic growth: The so-what question. *Journal of Economic Growth* 8: 47–71.

Henderson, J., and H.G. Wang. 2005. Urbanization and city growth. *Journal of Economic Geography* 5: 23–42.

Kolko, J. 1999. *Can I get some service here? Information technology service industries, and the future of cities*. Cambridge, MA: Mimeo/Harvard University.

Krugman, P. 1991. Increasing returns and economic geography. *Journal of Political Economy* 99: 483–499.

Lee, T.C. 1997. *Industry decentralization and regional specialization in Korean manufacturing*. Ph.D. thesis. Providence, RI: Brown University.

Lewis, W. 1954. Economic development with unlimited supplies of labor. *Manchester School of Economic and Social Studies* 22: 139–191.

Puga, D. 1999. The rise and fall of regional inequalities. *European Economic Review* 43: 303–334.

Renaud, B. 1981. *National urbanization policy in developing countries*. New York: Oxford University Press.

Williamson, J. 1965. Regional inequality and the process of national development. *Economic Development and Cultural Change* 13(4): 3–45.

Clapham, John Harold (1873–1946)

Phyllis Deane

Keywords

Clapham, J. H.; Economic development; Economic history; Marshall, A

JEL Classifications

B31

Sir John Clapham, who became in 1928 the first professor of economic history in the University of Cambridge, was born in Lancashire, the son of a prosperous jeweller. From the Cambridge boarding school (Leys) to which he was sent at the age of 14, he went up to King's College in 1892 to read history at a time when Acton, Maitland and Cunningham dominated the history school. It was as a graduate student at King's, researching into the French Revolution, that he attracted the attention of Alfred Marshall, who characteristically set about pressuring the promising young historian to devote his research efforts to filling the gaps in modern English economic

history. There is an oft-quoted letter which Marshall wrote in 1897 to Acton saying:

> I feel that the absence of any tolerable account of the economic development of England during the last century and a half is a ... grievous hindrance to the right understanding of the problems of our time ... but till recently the man for the work had not yet appeared. But now I think the man is in sight. Clapham has more analytic faculty than any thorough historian whom I have ever taught: his future work is I think still uncertain: a little force would I think turn him this way or that. If you could turn him towards XVIII or XIX century economic history, economists would ever be grateful to you.

Unfortunately Marshall did not live to read Clapham's massive, three-volume *Economic History of Modern Britain,* the first volume of which appeared in 1926 (dedicated to Marshall and his old enemy William Cunningham), and the last in 1938. No doubt he approved of the scholarly monograph on *The Woollen and Worsted Industries* (1907), written when young Clapham was professor of economics at the University of Leeds – an appointment in which it is hard not to suspect that Marshall's influence was decisive. Nevertheless, when Clapham returned to a King's fellowship in 1908, he resumed his researches in French political history and joined his fellow historians in criticizing the new Economics Tripos for being far too theoretical. It was not until after the First World War (during which he served in the Board of Trade and gained first-hand experience of the process of economic decision-making as a member of the Cabinet Committee on Priorities) that he in effect rejoined the path that Marshall had pointed out to him. His *Economic Development of France and Germany* (1921) was the first modern study in comparative economic development, but typically it involved juxtaposing his detailed analyses of two differing experiences of development, rather than relating them to a general theory of economic development, or even generalizing from these case histories.

The truth is that Clapham had no interest in theoretical economics except in so far as it supplied concepts and categories that would permit him to classify and analyse the empirical detail of economic history. He was repelled by the blatant unrealism of orthodox theorizing. His famous article 'Of Empty Economic Boxes', published in the September 1922 *Economic Journal*, accused the theorists of operating with concepts which were empty and irrelevant. 'I think a great deal of harm has been done', he complained, 'through omission to make clear that the Laws of Return have never been attached to specific industries: that we do not, for instance, this moment *know* under what conditions of returns coals or boots are being produced'. But his complaints fell on deaf ears. The interwar theorists saw no point in relating the strategic concepts of their models to real-world constructs and were agreed that, as Keynes put in, Clapham was 'barking up the wrong tree'.

What Clapham had learned from Marshall was that economics is the study of mutually interacting quantities and that it was the function of an economic historian to put the key quantitative questions to the historical record – for example, how large? how long? how often? how representative? – when spelling out the chains of cause and effect linking economic events. He made it his business to demolish, or qualify, facile generalizations that did not stand up to the available statistical evidence; for example, the Malthusian law of population, or the Marxian predictions of the pauperization of the masses. Though alive to the defects of historical statistics, he was bold enough to make the best of them, 'to offer dimensions, in place of blurred masses of unspecified size' and to analyse the bare aggregates into their strategic components. His training as a historian, however, kept a balance between quantitative and qualitative data, and his large-scale study of the economic development of modern Britain was diversified and illuminated by a continuous stream of vivid factual detail. His last book, *The Bank of England: A History, 1694–1914* (1944), commissioned by the Bank to commemorate its 250th anniversary, gave him access to the voluminous manuscript records of the first central bank. Writing its history and setting its operations and policies within its political and economic context was a task which by training and interests he was peculiarly well-equipped to perform. His intellectual energy seemed enhanced rather

than diminished by his retirement from the Cambridge chair, and his sudden death in 1946 cut short a research programme which was still in full swing.

Selected Works

1907. *The woollen and worsted industries*. London: Methuen.
1921. *The economic development of France and Germany, 1815–1914*. Cambridge: Cambridge University Press.
1922. Of empty economic boxes. *Economic Journal* 32: 305–314.
1926–38. *An economic history of modern Britain*. 3 vols. Cambridge: Cambridge University Press. Vol. 1, *The early railway age 1820–1850* (1926); vol. 2, *Free trade and steel* (1932); vol. 3, *Machines and national rivalries (1887–1914) with an epilogue (1914–1929)* (1938).
1944. *The bank of England: A history, 1694–1944*. 2 vols. Cambridge: Cambridge University Press.

Clark, Colin Grant (1905–1989)

H. W. Arndt

Keywords

Aggregate demand; Agriculture and economic development; Clark, C. G.; Economic growth; Multiplier; National income; Population growth

JEL Classifications

B31

Colin Clark, one of the most fertile minds in 20th-century applied economics, was born in London. After graduating in chemistry at Oxford University in 1924, he worked as assistant to W.H. Beveridge, Allyn Young and A.M. Carr-Saunders, stood unsuccessfully as a Labour candidate in the May 1929 general election, then joined the staff of the Economic Advisory Council, recently set up by Ramsay MacDonald, of which Keynes was a member. In 1931, rather than agree to write a protectionist manifesto for MacDonald, he accepted an appointment as lecturer in statistics at Cambridge, where he remained until, in 1937, he went to Melbourne University, initially as visiting lecturer. In Australia he occupied government posts, chiefly as economic adviser to the state government of Queensland, until 1952. After spells as visiting professor at the University of Chicago and as Director of the Oxford Institute of Agricultural Economics, he returned to Australia in 1968. He remained active as a research consultant at the University of Queensland.

In the first decade of an astonishingly prolific half-century of research and writing, Colin Clark established himself as one of the pioneers of national income estimates. He greatly improved existing estimates for the United Kingdom, and later for Australia and the Soviet Union, and in so doing made methodological contributions so fundamental that he has justly been described as co-author, with Simon Kuznets, of the 'statistical revolution' that accompanied the revolution in macroeconomics of the 1930s. He was the first to use the gross national product (GNP) and to present estimates in the framework of the main components of aggregate demand ($C+I+G$); he made some of the earliest estimates of Keynes's multiplier and, in an article published in 1937, one of the first international comparisons of the purchasing power of national currencies and thus of real national product. These were carried further in his monumental *Conditions of Economic Progress* (1940), which was important chiefly because it signalled the revival of interest among the profession in secular economic growth and development but which also supplied the first substantial statistical evidence of the gulf in living standards between rich and poor countries (the 'Gap') and developed the thesis that, in the course of economic growth, the occupational structure

shifts from primary to secondary and tertiary industries. During the Second World War, in *The Economics of 1960* (1942), Clark made one of the first ambitious attempts at a macroeconomic model of the world economy.

Recognized also as one of the 'Pioneers in Development', Colin Clark made significant contributions to empirical study of the relations between food supply and population growth, the economics of irrigation and subsistence agriculture, of determinants of economic growth and of productivity in agriculture in developing countries. At the same time, he was a gadfly in the political economy of developed countries, arguing against growthmanship, against high taxation and against welfarism long before it became fashionable to do so.

See Also

► Stone, John Richard Nicholas (1913–1991)

Selected Works

1932. *The national income 1924–31*. London: Macmillan.
1937. *National income and outlay*. London: Macmillan.
1939. *A critique of Russian statistics*. London: Macmillan.
1940. *Conditions of economic progress*. London: Macmillan. Revised ed, 1957.
1942. *The economics of 1960*. London: Macmillan.
1961. *Growthmanship*. London: Institute of Economic Affairs.
1964 (With M.R. Haswell.). *Economics of subsistence agriculture*. London: Macmillan.
1967. *Population growth and land use*. London: Macmillan. Revised ed, 1977.
1982. *Regional and urban location*. St Lucia: University of Queensland.
1984. Development economics: The early years. In *Pioneers in development*, ed. G.M. Meier and D. Seers. New York: Oxford University Press.

Clark, John Bates (1847–1938)

Donald Dewey

Keywords

Adams, H. C.; American Economic Association; Antitrust policy; Arbitration; Böhm-Bawerk, E. von; Capital accumulation; Capital measurement; Capital theory wages fund; Clark, J. B.; Clark, J. M.; Ely, R. T.; Factor shares; George, H.; German historical school; Jevons, W. S.; Johnson, A. S.; Knies, K. G. A.; Marginal utility principle; Natural and actual values; Neoclassical theory of distribution; Trade unions; Rent; Social Darwinism; Sraffa, P.; Sumner, W. G.; Universal measure of value; Veblen, T

JEL Classifications

B31

John Bates Clark, the first American economist to deserve and gain an international reputation, was born at Providence, Rhode Island, on 26 January 1847 into a modestly prosperous merchant family. His father's struggle with tuberculosis prompted a move to Minneapolis in search of a better climate and later required Clark to discontinue his studies at Amherst (he had transferred from Brown after two years) in order to run the family business. The business involved selling a line of ploughs to receptive but credit-needy country storekeepers throughout Minnesota. Following his father's death, the business was sold at a profit and Clark returned to Amherst, graduating with highest honours in 1872.

Clark's New England forebears had included many Congregational ministers and he seriously considered entering the Yale Divinity School. (He remained a communicant throughout his life and saw one son enter the ministry.) But encouraged by President Julius Seelye of Amherst, who had taught him political economy out of Amasa Walker's textbook, he chose instead the high-risk

course of an academic career in a country still without universities. After Amherst, he went abroad, enrolling for two years at Heidelberg and six months at Zurich.

While Clark has left no detailed account of his European studies, his early work indicates that he was much influenced by the German Historical School, and especially by the lectures of Karl Knies. Whether the influence was for good or ill is not clear. It probably slowed his development as a theorist. (His formulation of the marginal utility principle was worked out before he had heard of Jevons.) But it also taught him that an economist needed a far more professional training than that provided by the thin textbook gruel offered in the American colleges of the day. Clark was one of three young 'Germans' (the other two being Richard Ely and Henry Carter Adams) who, at a meeting of the American Historical Society at Saratoga in 1885, issued the call that led to the formation of the American Economic Association. Their plainly avowed purpose was to encourage German-style empirical research and give a sympathetic hearing to the critics of laissez faire. The dogmatic social Darwinism of William Graham Sumner epitomized all that they disliked in American economics. Clark became the third president of the new group and his diplomacy and moderation are credited with making it more acceptable to the country's older economists, most of whom eventually joined (but not Sumner).

Shortly after going to his first professorship at Carleton College in Northfield, Minnesota, in 1876, Clark was incapacitated for two years by an illness that, according to his son, John Maurice, permanently lowered his energy level. Whatever its nature – the family memorial to Clark provides no details – the illness seems only to have strengthened his determination and powers of organization. Following his recovery, Clark worked steadily and with a notable economy of effort until shortly before his death at the age of 91. Most of his contributions to economic theory, however, were worked out in the first 15 years of his career though the most polished formulations did not come until *The Distribution of Wealth* (1899). Clark's need to choose his projects carefully may explain why, despite his admiration for the work of historians and institutionalists, he never tried to emulate them. All of his life Clark remained a theorist who often wrote on issues of the day.

Clark first gained recognition with a series of articles in *The New Englander* that, with revisions, were published in 1886 as *The Philosophy of Wealth*. Clark's admirers have found this first book something of an embarrassment, and not without reason. It is a young Victorian's book, full of grand historical generalizations and the elevated expressions of sentiment that have long been out of fashion. Still, on close reading, it reveals the qualities that were to make him a major figure in the history of economics – a superb command of language (Böhm-Bawerk, who debated capital theory with Clark, claimed that his literary elegance gave him an unfair advantage), a willingness to take a position on controversial issues, and, above all, a remarkable talent for economic theory.

The collection contains a totally original and quite sophisticated statement of the principle of marginal utility ('effective utility' in Clark's vocabulary), a reasoned rejection of Malthusian pessimism, and many perceptive comments on the rise of labour unions, cartels, and corporations. Even the main outlines of Clark's treatment of capital and interest are discernible in the *Philosophy.*

Clark's intellectual distinction was fully revealed two years later with the publication of his monograph, *Capital and its Earnings* (1888a) which has a good claim to stand as the foundation stone of modern capital theory. While the distinction between labour and capital is still accepted (though even here Clark wavers), all other things including land that directly or indirectly enter into the production of consumer goods are treated as capital. The existence of interest is firmly placed in the productivity of capital. The creation of income as a concomitant of the destruction of individual capital goods is emphasized. The irrelevance of the 'period of production' of individual capital goods to anything of importance is shown and the fallacy underlying the wages fund doctrine is exposed.

Clark has been criticized for introducing the 'neoclassical fairy tale' into capital theory – the

notion that capital is some strange substance that, 'transmutes itself from one machine form into another like a restless reincarnating soul' (Samuelson 1962). While the neoclassical fairy tale has its limitations as a construct for understanding capital accumulation in the real world, Samuelson's jibe is off target. Clark's view of the production process is perfectly correct. Machines do 'transmute' themselves into other machines in the course of wearing out.

A more serious challenge to capital theory in the Clark tradition goes back to Böhm-Bawerk. If there is such a thing as a quantity of capital 'embodied' at any given moment in a set of heterogeneous specialized capital goods, what is its unit of measure? Unlike Irving Fisher, Clark faced the question squarely and attempted an answer. Unfortunately, the effort led him to bring forth his 'universal measure of value' – the product of a strange and nearly unintelligible fusion of utility analysis and the labour theory of value. While Clark was inordinately proud of his measure (and credited its inspiration to some lectures of Knies) it quickly found a merciful oblivion.

Later writers in the Clark tradition – or, at any rate, those who have felt the need for an impeccably consistent set of assumptions – have curbed their ambitions and been content to solve (or evade) the measurement problem by positing a surrogate production function where all capital goods are moulded from some homogeneous putty-like substance. The limit case in the Clark tradition is the 'Crusonia plant' named by Frank Knight but first suggested by W.S. Jevons's 'whole produce'. It supplies all human wants and, in the absence of consumption, grows at a constant geometric rate. Here the quantity of capital can be found either by measuring Crusonia directly or by dividing the plant's yield (income) in perpetuity by its natural growth rate, that is, the marginal (and average) productivity of investment.

Whether one prefers capital theory in Clark's tradition to its principal rival – capital theory in the Sraffa tradition – is ultimately a matter of personal taste. Both employ simplifications that take one far from reality. However, notwithstanding the measurement conundrum, to date capital theory in the Clark tradition has provided the basis for virtually all empirical work on wealth and income. This is not surprising. To statisticians, measuring changes in the quantity of capital (which they rename the real value of the stock of capital assets) is just another index number problem.

Very early in his career Clark began to work on the problem of factor shares (possibly because of his interest in Henry George) and concluded that the treatment of land rent as a surplus whose size is not determined by marginal productivity was gross error. The most complete statement of his views on distribution is in *The Distribution of Wealth* (1899) which drew heavily on his earlier articles and monographs. Despite its flaws (which include the universal measure of value) the *Distribution* is a remarkable book and, by any reasonable test, a landmark treatise in the development of economics.

The *Distribution* represents an advance on the prior art in two important respects. It offers a discussion of the relation of statics to dynamics – the terms were introduced into economics by Clark – superior to that of previous treatments. And it offers, for the first time, a complete and lucid exposition of the neoclassical theory of distribution. The *Distribution* also brought Clark's views on capital to a much wider audience.

Clark was as conscious of the rapid pace of economic change as any German or American institutionalist of his day, but he stressed that, at any given moment, there are 'natural' values in the marketplace and permanent pressures pushing actual values toward them.

> Reduce society to a stationary state, let industry go on with entire freedom, make labor and capital absolutely mobile – as free to move from employment to employment as they are supposed to be in the theoretical world that figures in Ricardo's studies – and you will have a regime of natural values. These are the values about which rates are forever fluctuating in the shops of commercial cities. You will also have a regime of natural wages and interest; and these are the standards about which the rates of pay for labor and capital are always hovering in actual mills, fields, mines, etc.

Only by a careful separation and delineation of static and dynamic forces, Clark believed, can the

process of price formation in real-world markets be understood. His methodology is not as formal and austere as F.H. Knight's in *Risk*, *Uncertainty*, *and Profit* (1921), but it is essentially the same. (In the version of Knight's doctoral dissertation accepted at Cornell in 1916 his intellectual debts to Clark are gratefully and fully acknowledged; for reasons unknown, almost all of the favourable references to Clark are omitted in the rewritten version published five years later).

To demonstrate that, in the static state, payments to the factors exhaust the product when each receives its marginal product, Clark devised a set of diagrams to show that, in a two-factor model, what is viewed as rent and what is viewed as a factor payment is a matter of perspective. One becomes the other by interchanging the fixed and variable factors in the diagrams. Clark's treatment of rent has been followed by an admiring Paul Samuelson in all of the many editions of his *Economics*.

Clark's approach to distribution is set forth in 'words and pictures' (his mathematical training did not include calculus) and so lacks the precision of the versions of Wicksell and Wicksteed. But, being more accessible to student readers, it was Clark's treatment that first gained widespread attention for the neoclassical theory of distribution.

Clark has often been reproved for implying both that factor payments ought to be according to marginal productivity and that in a real-world market economy most factor payments do closely approximate marginal productivity (see, for example, Stigler 1941). A reading of the *Distribution* without reference to Clark's other writings would indicate that he did hold these views. Certainly his advocacy of compulsory arbitration to end long labour disputes assumed that economic justice consisted in giving striking workers the wages prevailing in comparable employments elsewhere. However, a brilliant essay, 'The Theory of Economic Progress' (1896), leaves no doubt that he placed a far higher value on economic growth than on short-run justice or efficiency.

Well before Schumpeter, Clark wrote:

> The picture of a stationary state presented by John Stuart Mill as the goal of competitive industry is the one thing needed to complete the impression of dismalness made by the political economy of the early period. A state could not be so good that that lack of progress would not blight it; nor could it be so bad that the fact of progress would not redeem it. . . . The decisive test of an economic system is the rate and direction of movement.

Clark was a leading participant in the trust controversy that occupied American politics in the 30 years before the First World War. His moral seriousness and literary ability (and, one suspects, his ability to meet deadlines) made him a favourite of magazine editors – he once described himself as 'writing my trust article again'. Like all economists of that era he had to think through his attitude toward the many large firms with large market shares that had so suddenly appeared.

As recorded in the *Philosophy of Wealth*, Clark's first reaction to the American business scene on returning from Germany was one of fascinated revulsion joined to an expression of hope that businessmen could be led to behave in more acceptable ways by pressures from labour unions, Church, and State. As the years passed, his views of commerce became much more favourable and his policy recommendations more worldly and specific. He early pointed out that the conduct of most so-called trusts was influenced by the fear of entry and he never depreciated the efficiency gains made possible by large-scale production. At first he urged only a modest amount of government intervention as in, *The Control of Trusts*: *An Argument in Favor of Curbing the Power of Monopoly by a Natural Method* (1901). Clark's 'natural method' was little more than the competition of the marketplace purged of its 'destructive' ingredients plus government regulation of railroad rates to prevent unjustified differentials. A much expanded version of *The Control of Trusts*, with John Maurice Clark, his son, as co-author and the subtitle omitted, appeared in 1912. The revisions were mostly the work of the son and contain a virtual blueprint for an antitrust policy. The Clayton and Federal Trade Commission Acts of 1914 which followed shortly received their enthusiastic approval.

By his writing Clark did more than any other economist to confer intellectual respectability on

an antitrust policy that had had its origins in the populist discontent that produced the Sherman Act. In retrospect, this may seem to have been a dubious achievement. But in Clark's favour it can be said that he was dealing with new and difficult issues and approached them with more objectivity than most of his contemporaries, for example, W.Z. Ripley and F.A. Fetter.

Clark's life as a teacher was at Carleton, Smith, Amherst, and from 1895 to 1923 at Columbia. At Carleton his kindness helped Thorstein Veblen (a thoroughly unpopular undergraduate in that church college) to find his way. At Columbia it helped Alvin Johnson to gain the income needed to complete his doctoral programme. His encouragement led F.H. Giddings to leave provincial journalism for a seminal career in sociology. He was, of course, the omnipresent influence in the life of John Maurice Clark, who succeeded to his chair at Columbia. Still, Clark's direct influence through the classroom seems to have been surprisingly limited. His quiet and self-sufficient personality did not require disciples and his probing but loosely organized lectures appealed only to very able students. Then too, Clark was a theorist in an era when, in the United States, institutional economics, not theory, was the height of academic fashion.

From 1911 onward Clark's great concern became the contribution that social scientists could make to ending war. When the Carnegie Endowment for International Peace was formed in 1910, he became the first director of its economics and history section serving until 1923. There he took the initiative in obtaining support for the studies that became the *Social and Economic History of the World War*. The general editor was his friend and Columbia colleague in history, James T. Shotwell. The Carnegie *History* ultimately ran to over a hundred volumes and still stands as the most ambitious research project in the social sciences ever undertaken by a private foundation. Unfortunately, its initial promise was never realized. Shotwell sought to organize the Carnegie *History* on the strange principle that an accounting of the great war was too important to be left to historians. As a result, while the series contains a few memorable studies, for example, J.M. Clark, *The Costs of the World War to the American People* (1931), it served mainly to preserve the recollections of wartime ministers and civil servants that would otherwise have been lost. J.M. Keynes disdainfully withdrew from the *History* in the planning stage.

Clark's work for peace continued to the end of his life. His last small book was a moving plea for collective action to deter aggression, *A Tender of Peace*: *The Terms on Which Civilized Nations Can*, *if They Will*, *Avoid Warfare* (1935). Clark died in New York City on 21 March 1938.

An abundance of honours came to him in his lifetime both in the United States and abroad. They were all deserved.

See Also

- ▶ Clark, John Maurice (1884–1963)
- ▶ Fisher, Irving (1867–1947)
- ▶ Marginal Productivity Theory
- ▶ 'Neoclassical'

Selected Works

1886. *The philosophy of wealth: Economic principles newly formulated*. Boston: Ginn & Co.

1888a. *Capital and its earnings*. Baltimore: American Economic Association.

1888b (With F.H. Giddings). *The modern distributive process*. Boston: Ginn & Co.

1893. The ultimate standard of value. *The Yale Review* 1: 252–274.

1896. The theory of economic progress. *American Economic Association*: *Economic Studies* 1: 1–22.

1899. *The distribution of wealth: A theory of wages, interest and profits*. New York: The Macmillan Co.

1901. *The control of trusts: An argument in favor of curbing the power of monopoly by a natural method*. New York: Macmillan.

1904. *The problem of monopoly: A study of a grave danger and of the natural mode of averting it.* New York: Columbia University Press.

1907. *The essentials of economic theory: As applied to modern problems of industry and public policy.* New York: Macmillan.
1912 (With J.M. Clark). *The control of trusts*. New York: Macmillan.
1914. *Social justice without socialism*. Boston: Houghton Mifflin.
1935. *A tender of peace: The terms on which civilized nations can, if they will, avoid warfare.* New York: Columbia University Press.
A nearly complete listing of Clark's publications is in A Bibliography of the Faculty of Political Science, Columbia University, 1880–1930, New York: Columbia.
University Press, 1931; also in Economic Essays Contributed in Honor of John Bates Clark, ed. J.H. Hollander, New York: Macmillan, 1927.

Bibliography

Böhm-Bawerk, E. 1906. Capital and interest once more. *Quarterly Journal of Economics* 21 (1–21): 247–282.
Clark, J.M. 1931. *The costs of the world war to the American people*. New Haven: Yale University Press.
John Bates Clark. 1938. *A memorial volume prepared by his children*. New York (privately printed).
Knight, F.H. 1921. *Risk, uncertainty, and profit*. Boston: Houghton Mifflin.
Samuelson, P. 1962. Parable and realism in capital theory: The surrogate production function. *Review of Economic Studies* 29: 193–206.
Stigler, G. 1941. *Production and distribution theories: The formative period*. New York: Macmillan.

Clark, John Maurice (1884–1963)

Warren J. Samuels

Keywords

American Economic Association; Business cycles; Clark, J. B.; Clark, J. M.; Commons, J. R.; Dewey, J.; Institutional economics; Overhead costs; Social economics; Veblen, T.; Workable competition

JEL Classifications

B31

Clark was born on 30 November 1884 in Northampton, Massachusetts, and died on 27 June 1963 in Westport, Connecticut. Educated at Amherst College and Columbia University (Ph.D., 1910), he taught at Colorado College (1908–10), Amherst (1910–15), University of Chicago (1915–26) and Columbia University (1926–52), where he succeeded his father, John Bates Clark. He was president of the American Economic Association in 1935 and received its Francis A. Walker Medal in 1952. His dissertation, 'Standards of Reasonableness in Local Freight Discrimination', was written under the supervision of his father. He was associated with the National Bureau of Economic Research, the National Resources Planning Board, the Twentieth Century Fund, the Attorney General's National Committee to Study the Anti-Trust Laws, and other organizations.

Clark worked within both orthodox and heterodox economics, making important contributions to microeconomics, macroeconomics and institutional, or social, economics. Eclectic and open-minded, he was critical of the apologetic uses of economic theory, particularly of the drawing of narrow and misleading welfare implications. He emphasized the limits of economics as a science.

Clark's contributions within conventional theory dealt principally with economic dynamics. He developed and stressed the implications of overhead, fixed costs in capital intensive industry for competitive structure, business pricing policy, and economic stability. He was the principal of several discoverers of the acceleration principle, with its important implications for instability. His career-long concern with competitive structure and behaviour led to his formulation of the concept of 'workable competition', with a stress on potential competition and intercommodity substitution. The major result of his equally long work in macroeconomics was an exploration of the strategic factors in business cycles which effectively summarized, in a general theoretical context, the state of empirical knowledge at the time. He also

wrote extensively on railroad and public utility rates, basing-point pricing, economic planning, the economics of war and of peacetime conversion, wage-price (cost-push inflation) theory and policy, and related topics.

Clark departed from the conventional mainstream in his social economics, which was akin to the institutional economics of John R. Commons and Wesley C. Mitchell and which reflected the influence of Thorstein Veblen and John Dewey. Clark's work on the social control of business and the theory of regulation explored the fundamental legal-economic nexus of society in a non-ideological manner stressing the substance and inexorable presence of formal (legal) and informal controls in an economic system, even in a pluralistic and voluntaristic economy, controls typically obscured in conventional analysis of markets. Law was important to the structure of freedom, not something solely antagonistic to freedom. His work in welfare economics emphasized the role of institutions, the necessity of psychological realism, and the inexorable role of moral or ethical values. His concern with the costs of labour that are registered in neither the market nor by industry presaged later institutional work on externalities and social costs.

Selected Works

1912. (With J.B. Clark.) *The control of trusts.* New York: Macmillan.
1923. *Studies in the economics of overhead costs.* Chicago: University of Chicago Press.
1926. *Social control of business.* New York: McGraw-Hill.
1934. *Strategic factors in business cycles.* New York: H. Wolff for the National Bureau of Economic Research.
1936. *Preface to social economics.* New York: Farrar and Rinehart.
1948. *Alternative to serfdom.* New York: Knopf.
1949. *Guideposts in time of change.* New York: Harper.
1957. *Economic institutions and human welfare.* New York: Knopf.
1961. *Competition as a dynamic process.* Washington, DC: Brookings Institution.

Class

David B. Grusky

Abstract

The structure of inequality has historically been represented with an income paradigm that treats well-being as adequately indexed by income alone. By contrast, the class-analytic tradition treats inequality as fundamentally multidimensional, with such variables as health, education and social relations all deemed important nonincome constituents of well-being. These variables may assume a class-based form in which social groups within the division of labour define characteristic constellations of scores. The class model is further supported in so far as class membership has true causal effects on behaviours that are not reducible to the effects of income or other correlates of class.

Keywords

Class; Compensating differentials; Human development index; Identity; Inequality, multidimensional; Labour markets; Marx, K.; Redistribution of income and wealth; Sen, A.; Socio-economic index; Underclass

JEL Classifications

B31

The labour market of contemporary societies is rife with various types of 'classes' that impede the free flow of labour by restricting entry to those who have the requisite degrees, certificates, memberships or capital. These classes take the form, for example, of occupations (such as economist, carpenter), aggregates of occupations (such as manager, farmer), or groups that represent competing factors of production (such as worker, capitalist). Although such classes are ubiquitous in contemporary labour markets, their effects on labour market processes are not

always incorporated into formal economic models. The main type of class to which attention has historically been paid is that of industry. The bifurcation of labour markets into industry classes, while clearly a relevant and well-developed topic in the literature, is not covered here. For purely historical reasons, the term 'class' has been reserved for non-industrial forms of bifurcation, a usage that is adopted in the following discussion as well.

The descriptive rationale for a class model is usefully introduced in the context of a multidimensional representation of inequality. This representation, which is presented below, makes it possible to motivate the class concept, to consider how classes may be empirically revealed, and to assess whether the class concept is needed to represent the structure of labour markets.

The Clustering Rationale

It has become increasingly fashionable to claim that inequality is multidimensional, that income inequality is accordingly only one of many important forms of inequality, and that income redistribution in and of itself would not eliminate inequality (see, for example, Sen 2006). If this line of argument is taken seriously, an obvious prescription is to examine separately each of the many variables that constitute the multidimensional space of interest. For example, one might usefully distinguish between the eight forms of inequality listed in Table 1, each such form pertaining to a type of good that is intrinsically valuable (as well as possibly an investment). The multidimensional space formed by these variables may be labelled the 'inequality space'. The social location of an individual within this

Class, Table 1 Types of valued goods and examples of advantaged and disadvantaged groups

Valued goods		Examples	
Type	Example	Advantaged	Disadvantaged
1. Economic	Wealth	Billionaire	Bankrupt worker
	Income	Professional	Laborer
	Ownership	Capitalist	Employee
2. Power	Political power	Prime minister	Disenfranchised person
	Workplace authority	Manager	Subordinate worker
	Household authority	'Head of household'	Child
3. Cultural	Knowledge	Intelligentsia	Uneducated
	Popular culture	Movie star	High-culture 'elitist'
	'Good' manners	Aristocracy	Commoner
4. Social	Social clubs	Country-club member	Non-member
	Workplace associations	Union member	Non-member
	Informal networks	Washington 'A list'	Social unknown
5. Honorific	Occupational	Judge	Garbage collector
	Religious	Saint	Excommunicate
	Merit-based	Nobel Prize winner	Non-winner
6. Civil	Right to work	Citizen	Illegal immigrant
	Due process	Citizen	Suspected terrorist
	Franchise	Citizen	Felon
7. Human	On-the-job	Experienced worker	Inexperienced worker
	General schooling	College graduate	High-school dropout
	Vocational training	Law-school graduate	Unskilled worker
8. Physical	Mortality	Person with long life	A 'premature' death
	Physical disease	Healthy person	Person with AIDS, asthma
	Mental health	Healthy person	Depressed, alienated

inequality space can then be characterized by specifying her or his constellation of scores on each of the eight types of variables in this table.

At least implicitly, scholars of inequality long ago adopted precisely such a multidimensionalist approach, as revealed by the burgeoning research literatures that monitor not just income inequality but also inequality of health, social networks, education, computer usage and all manner of other valued goods. This line of research typically takes the form of an exposé of the extent to which seemingly basic human 'entitlements', such as living outside of prison, being gainfully employed, freely participating in digital culture, or living a reasonably long and healthy life, are unequally distributed in ways that may amplify or somehow complement well-known differentials of income or earnings.

Does the inequality space take on a simpler form than might be implied by the convention of analysing each of these variables separately and independently? Two possible simplifications may be considered here. First, scholars have frequently combined scores on the underlying variables to form indices, with sociologists often combining education and income into a socio-economic index (for example, Hauser and Warren 2001) and development economists often combining measures of health, income, education and literacy into a ' Human Development' index (for example, UNDP 2005). There is, however, growing concern that such standard multidimensional scales are excessively abstract and fail to capture the social organization of inequality, especially the emergence of social networks, norms, and adaptive preferences or tastes among individuals in similar life situations and circumstances. The socio-economic scale, for example, is a purely statistical tool that groups together individuals of similar income or education levels without any consideration of whether these individuals associate with one another or are comembers of some real group, such as a union or occupation.

This critique motivates a second, class-based approach to understanding the structure of the inequality space. The class model is defensible insofar as (*a*) individuals tend to cluster into a relatively small number of characteristic combinations or packages of scores on the underlying variables, and (*b*) the clusters are defined by such structural locations as detailed occupations (doctor, secretary, plumber), aggregates of detailed occupations (professional, manager, clerk, craft worker, labourer, farmer), or other types of 'big classes' (for example, capitalist, worker). These clusters generate a labour market that, instead of being a seamless distribution of incomes, is a lumpy entity with deeply institutionalized groups that constitute pre-packed combinations of valued goods.

The class of craft workers, for example, has historically comprised individuals with moderate educational investments (secondary-school credentials), considerable occupation-specific investments in human capital (vocational or on-the-job training), average income, relatively high job security, middling social honour and prestige, quite limited authority and autonomy, and comparatively good health outcomes (by virtue of union-sponsored health benefits and regulation of working conditions). By contrast, the underclass is characterized by a rather different package of scores, one that combines minimal educational investments, limited opportunities for on-the-job training, intermittent labour force participation, low income, virtually no opportunities for authority or autonomy on the job (during brief bouts of employment), relatively poor health (by virtue of lifestyle choices and inadequate health care), and much social denigration and exclusion. The other classes appearing in class schemes (such as professionals, managers, clerks, labourers, farmers) may likewise be understood as particular combinations of scores on the variables of interest.

In a class-based society, the inequality space will accordingly have relatively low dimensionality, a dimensionality no more or less than the number of classes. This understanding of the class principle implies that the variables constituting the inequality space must be independent of one another *within* each class. If the independence assumption begins to break down within a postulated class, we can then speak of 'subclasses' forming by virtue of developing their own distinguishable packages of scores. It is useful in this

context to distinguish between a big-class regime in which the dimensionality of the inequality space is small and a micro-class regime in which the dimensionality of the inequality space is large. Although Marx (1894) argued that the inequality space in the early industrial period was becoming increasingly consistent with a two-class solution (in which privileged capitalists were juxtaposed to disadvantaged workers), some contemporary class analysts emphasize, to the contrary, that the forces of market differentiation have generated a micro-class regime in which the independence assumption holds not at the big-class level but only within quite detailed occupations (for example, Weeden and Grusky 2005). There is much ongoing debate among inequality scholars on the dimensionality of the contemporary inequality space and, in particular, on whether the dimensionality of that space has been increasing or diminishing.

The foregoing implies that one may usefully distinguish between big-class regimes with few classes and micro-class regimes with many classes. Additionally, one might distinguish inequality regimes not on the basis of how many classes there are but on the basis of how the classes differ from one another. In a purely 'vertical' class system, one can readily order classes on a single scale from 'low' to 'high', with low classes being systematically disadvantaged on all variables and high classes being systematically advantaged on all variables. This organization of the inequality space implies a stark form of inequality in which privilege on one dimension implies very reliably privilege on another. Alternatively, a class system that is (partly) horizontal will embody compensating forms of advantage and disadvantage, meaning that at least some classes are formed by combining high values on one dimension with low values on another. There is, again, much debate among class analysts as to whether the inequality space is becoming more or less vertically organized.

It is of course possible that the inequality space is organized in ways that are largely inconsistent with the class principle. Two types of non-class solutions, as reviewed below, may be usefully distinguished.

Extreme Disorganization

First, one can imagine an inequality space in which the underlying variables don't covary at all, hence yielding a one-class solution or, equivalently, a non-class regime. To be sure, there would be much inequality under this hypothetical constellation of data, yet it would take a uniquely structureless form in which the independence assumption holds throughout the inequality space, not just within a given class. It is unlikely that such extreme disorganization would ever be realized, but some postmodernists (for example, Pakulski 2005) have argued that we are moving gradually toward this form. If they are correct, it means that the growth in income inequality is at least counterbalanced by a decline in the association between income and other valued goods. As with the horizontal class regime described above, here again we have a form of inequality that embodies much in the way of compensating differentials, although such differentials are not in this case packaged into institutionalized classes.

Individuals as Classes

The second main type of non-class solution arises when the variables constituting the inequality space are related to one another in perfectly linear fashion. When the data are configured in this way, it is no longer possible to identify a set of classes within which independence holds, as the underlying inequality variables continue to covary with one another no matter how much one disaggregates. We are left with an extreme micro-class solution in which the data thin out to the point where each individual becomes a class unto himself or herself. This solution is consistent, for example, with the claim that income is a master variable, that it perfectly signals all other individual-level measures of inequality, and that no higher-level class organization therefore appears. Obviously, this ideal type would never be empirically realized in such extreme form, but it is nonetheless important to ask whether it comes closer to being realized in some societies or time periods than in others.

The 'Class Effect' Rationale

We have to this point represented the class principle as a hypothesis about the clustering of

observations in the inequality space. As an alternative motivation for the class hypothesis, it is sometimes claimed that classes are social contexts that affect attitudes, behaviour, and individual action of many kinds. When this motivation is adopted, classes are not typically construed as information-rich social containers that capture many life conditions of interest, but rather as analytic categories that single out a particular social context that is presumed to be very consequential in defining interests. Under such a formulation, a class analyst will therefore typically nominate a single variable (for example, authority, ownership) as especially useful in understanding the sources of social behaviour, with the class categories then defined so as to capture differences across workers on that underlying variable of presumed consequence. The Marxian model, for example, famously embodies the claim that classes are best defined in terms of employment status alone, with the rationale for this definition being that employment status putatively defines interests and hence attitudes and behaviour (Marx 1894). In contemporary labour markets, the class of employed workers is of course very heterogeneous, thus motivating class analysts to introduce further distinctions within that class that are presumed to be consequential in defining interests and action. There is no shortage of such elaborated class models (Wright 2005).

When a class model is motivated by presumed class effects, it is important to establish that such effects are indeed truly causal. If, for example, one finds that seeming differences in the politics of professionals, managers, craft workers and other social classes disappear when income is controlled, then presumably one can refer only to an income effect on politics, not a true class effect. Why might net effects of class be detected even with rigorous controls? In addressing this question, what must first be stressed is that, even when classes are *defined* in terms of a single analytic variable, the resulting classes are nonetheless often organic packages of conditions; and the constituents of these packages may combine and interact in ways that lead to an emergent logic of the situation. The underclass, for instance, may be understood as a combination of negative conditions (intermittent labour force participation, limited education, low income) that, taken together, engender a sense of futility, despondency, or learned helplessness that is more profound than what would be expected from a model that simply allows for independent effects of each constituent class condition. To be sure, a committed reductionist might counter that, when modelling behaviour, one merely needs to include the appropriate set of interactions between the constituent variables. In so far as classes define the relevant packages of interacting conditions, such an approach just becomes an unduly complicated way of sidestepping the reality of classes.

This emergent logic of the situation may well be undergirded by a class culture. At one extreme, class cultures may be understood as nothing more than 'rules of thumb' that encode optimizing behavioural responses to prevailing environmental conditions, rules that allow class members to forgo optimizing calculations themselves and rely instead on cultural prescriptions that provide reliable short cuts to the right decision. In this vein, Goldthorpe (2000) argues that working-class culture is disparaging of educational investments not because of some maladaptive oppositional culture but because such investments expose the working class, more so than other classes, to a real risk of downward mobility. Typically, working-class children lack insurance in the form of substantial family income or wealth, meaning that they cannot easily recover from an educational investment gone awry (in the form of dropping out); and those who nonetheless undertake such an investment therefore face the real possibility of substantial downward mobility. The emergence, then, of a working-class culture that regards educational investments as frivolous may be understood as encoding that conclusion and thus allowing working-class children to undertake optimizing behaviours without explicitly engaging in decision- tree calculations. The behaviours that a rule-of-thumb culture encourages are, then, deeply adaptive because they take into account the endowments and institutional realities that class situations encompass.

The foregoing example may be understood as one in which a class-specific culture instructs recipients about the best means for achieving ends that

are widely pursued by *all* classes. Indeed, the prior rule-of-thumb account assumes that members of the working class share the conventional interest in maximizing labour market outcomes, with their class-specific culture merely instructing them about the approach that is best pursued in achieving that conventional objective. At the other extreme, one finds class-analytic formulations that represent class cultures as more overarching world views, ones that instruct not merely about the proper means to achieve ends but additionally about the proper valuation of the ends themselves. For example, some class cultures (such as aristocratic ones) place an especially high valuation on leisure, with market work disparaged as 'common' or 'polluting'. This orientation presumably translates into a high reservation wage within the aristocratic class. Similarly, oppositional cultures within the underclass may be understood as world views that place an especially high valuation on preserving respect and dignity for class members, with of course the further prescription that these ends are best achieved by (*a*) withdrawing from and opposing conventional aspirations, (*b*) representing conventional mobility mechanisms (for example, higher education) as tailor-made for the middle class and, by contrast, unworkable for the underclass, and (*c*) pursuing dignity and respect through other means, most notably total withdrawal from and disparagement of mainstream pursuits. This is a culture, then, that gives respect and dignity an especially prominent place in the utility function and that further specifies how respect and dignity might be achieved.

Whatever the mechanism that underlies class cultures and class effects, the common assumption is that classes are meaningful social contexts, just as neighbourhoods are likewise understood within the 'neighbourhood effects' literature as meaningful social contexts. These contexts are expected in both cases to have causal effects that are not reducible to mere selective processes. Again, we have to stress that such a 'class effects' rationale for class models is best treated as a hypothesis, as there is little in the way of substantiating evidence at this point (cf. Weeden and Grusky 2005).

It is altogether possible that such class effects are weak or at least weakening. The relevant postmodernist position in this regard is that social class has lost much of the power it once had because (*a*) other cross-cutting social cleavages (such as race or gender) have squeezed out class-based identities and interests, (*b*) identity formation in the postmodern world is so atomized and individualized that all structural bases of social behaviour have become less relevant, (*c*) the institutions that once represented class interests (for example, political parties, unions) have developed into new forms that are less class-based, or (*d*) the forces of the market work to gradually eliminate pockets of rent-generating social action. Regardless of the particular form of the argument, the expectation in all cases is that emergent effects of classes have, during the last several decades, become less prominent.

Conclusions

It should by now be clear that sociologists operating within the class-analytic tradition have adopted very strong assumptions about how inequality and poverty are structured. As was noted, the class concept may be motivated in two main ways, by claiming either that the inequality space has a (low) dimensionality equalling the number of social classes, or that the class locations of individuals have a true causal effect on behaviours or attitudes of interest. The foregoing claims have been unstated articles of faith among class analysts in particular and sociologists more generally. In this sense, class analysts have behaved rather like stereotypical economists, the latter frequently being criticized (and parodied) for their willingness to assume almost anything provided that it leads to an elegant model.

This critique of class analysis is, however, increasingly less justifiable. Indeed, the class-analytic status quo has come under much criticism of late, with many scholars now feeling sufficiently emboldened to argue that the concept of class should be abandoned altogether (for example, Kingston 2000; Pakulski 2005). Although the resulting debate has sometimes been unproductive, it has clearly precipitated an increasing interest in assessing the empirical foundations of class models.

See Also

- ▶ Economic Sociology
- ▶ Inequality (Global)
- ▶ Inequality (Measurement)
- ▶ Labour Economics
- ▶ Labour Market Institutions
- ▶ Poverty
- ▶ Social Status, Economics and

Bibliography

Goldthorpe, J. 2000. *On sociology: Numbers, narrative, and the integration of research and theory*. New York: Oxford University Press.

Hauser, R., and J. Warren. 2001. Socioeconomic indexes of occupational status: A review, update, and critique. In *Social stratification: Class, race, and gender in sociological perspective*, ed. D. Grusky. Boulder: Westview Press.

Kingston, P. 2000. *The classless society*. Stanford: Stanford University Press.

Marx, K. 1894. *Capital*, 3 vols. London: Lawrence and Wishart, 1972.

Pakulski, J. 2005. Foundations of a post-class analysis. In *Approaches to class analysis*, ed. E. Wright. Cambridge: Cambridge University Press.

Sen, A. 2006. Conceptualizing and measuring poverty. In *Inequality and poverty*, ed. D. Grusky and R. Kanbur. Stanford: Stanford University Press.

UNDP (United Nations Development Programme). 2005. *Human development report 2005*. New York: UNDP.

Weeden, K., and D. Grusky. 2005. The case for a new class map. *American Journal of Sociology* 111: 141–212.

Wright, E., ed. 2005. *Approaches to class analysis*. Cambridge: Cambridge University Press.

Classical Distribution Theories

Massimo Pivetti

Abstract

Classical distribution theories distinguish between that part of the annual product which is necessary for its reproduction (including necessary subsistence for workers and replacement of the means of production) and the remainder (the 'surplus'), and seek to explain the size of the surplus and its distribution among classes. They do not view the real wage rate and the rate of profit as determined by the relative scarcity of labour and capital; rather, one of the two distributive variables is explained independently from both the social product and the other distributive variable, and the other is determined as a residual.

Keywords

Classical distribution theories; Classical economics; Competition; German Historical School; Interest rates; Keynesian distribution theory; Marx, K. H.; Natural price; Real wage rate; Ricardo, D.; Smith, A.; Sraffa, P.; Surplus; Subsistence; Value theory; Wage–profit relationship

JEL Classifications

D6

The terms 'classical economists' and 'classical political economy' were first used by Marx, whose monumental survey of economic theory from the middle of the 17th century up to the early 1860s was contained in the manuscript written between January 1862 and July 1863 which the author called *Theorien über den Mehrwert*. Marx used the terms to describe 'the critical economists', 'the economic investigators . . . like the Physiocrats, Adam Smith and Ricardo' whose 'urge' was 'to grasp the inner connection of the phenomena'; he also referred to Ricardo as 'the last great representative' of classical political economy (Marx 1862–3, vol. 3, pp. 453, 500 and 502; 1873, p. 24).

Marx's description implies that not only authors like Senior, Bastiat, Wilhelm Roscher and John Elliot Cairnes are extraneous to classical political economy, but also such faithful Ricardians as James Mill, McCulloch and John Stuart Mill do not properly fit into it. This can only be understood if one bears in mind that the ranking of the various authors in *Theorien über den Mehrwert* is centred upon the nature of their contributions to the related subjects of distribution and value: the explanation of profit and the

formation of a normal or general rate of profit; the relation between wages and profits, the difficulties in the theory of value that arise in connection with the wage–profit relationship and the formation of a general rate of profit are the chief theoretical questions in the light of which the various authors are surveyed.

Thus a first discriminating factor in Marx's critical survey is provided by each author's attitude towards the main analytical difficulties: whether this or that author shows himself to be aware of their presence and tries to solve them, albeit at the cost of falling into further difficulties and contradictions, or rather tends to present the theory as a fully satisfactory body of propositions by denying the difficulties and 'immediately adapting the concrete to the abstract' (Marx 1862–3, vol. 3, p. 87). This factor explains why Marx is inclined to treat both Torrens (1815, 1821) and Malthus (in particular, 1827) as classical economists, while regarding James Mill as the beginner of the 'disintegration' of the Ricardian theory.

A second factor is the *weight* of the 'vulgar' element present in the contributions of the various authors – meaning by this the tendency to confine one's attention to the 'superficial appearance of the phenomena' *versus* 'the urge to grasp [their] inner connection'. As an important example of this factor one may refer to the increasing tendency, after Ricardo, to explain distribution by competition and 'the [changing] state of supply and demand' (J. Mill 1844, p. 42; see also J.S. Mill 1848, p. 337, and Cairnes 1874, pp. 168–74) – thereby gradually abandoning the classical conception according to which demand and supply can only determine the oscillations of distribution and prices either above or below their 'natural' values. A third discriminating factor is the 'vulgar' element represented by the mere apology for the existing state of affairs (Marx 1962–63, vol. 3, p. 168), or, as Cannan was later to put it, by the 'desire to strengthen the position of the capitalist against the labourer' (Cannan 1917, p. 206). Finally, a fourth factor may be indicated in the tendency to deny the existence of economic laws altogether, and to substitute shallow empiricism for theoretical analysis (think of the so-called Historical School of German political economy).

The theoretical approach to distribution and value 'of the old classical economists from Adam Smith to Ricardo has been submerged and forgotten since the advent of the "marginal" method' (Sraffa 1960, p. v). A contribution to this effect certainly came from the fact that *Theorien über den Mehrwert* remained largely unknown among economists. (It was only in the early 1950s that some sections of the 1905–10 Kautsky edition were translated into English, whilst the complete English translation from the edition based on the original manuscript was made in 1963–71.) In what follows, we shall take 'classical theory of distribution' to mean the main elements which can be regarded as characterizing the approach to the problem of the division of the national product among classes followed by the English classical economists from Adam Smith to David Ricardo, later by Karl Marx, and, more recently, by Piero Sraffa – this century's greatest exponent of the 'classical' approach to distribution.

The classical method of approaching the problem of distribution is based upon a distinction between two parts in the annual product of society: that part which is necessary for its reproduction (which includes the necessary subsistence of the workers employed in the economy) and that part which can be 'freely' disposed of by the society and which constitutes its 'net product' or 'surplus' – what remains of the social product after deducting the necessary subsistence of the workers and the replacement of the means of production. It is the aim of the classical theory to explain the circumstances governing the size of the surplus and its distribution among classes: 'To determine the laws which regulate this distribution, is', according to Ricardo, 'the principal problem in Political Economy' (Ricardo 1821, p. 5). In the course of his work he succeeded in 'getting rid of rent', so as to concentrate on the problem of the distribution between capitalists and workers; in what follows rent will be left entirely out of account – one may suppose that fertile lands abound – and the essential features of the surplus approach to distribution will be illustrated with

reference to the determination of wages and profits.

Contrary to the supply-and-demand approach, which has been the dominant method over the last hundred years, in the theoretical approach to distribution of the classical economists and of Marx, the real wage rate and the rate of profit are not symmetrically and simultaneously determined on the basis of the relative scarcity of labour and capital. Within the classical approach, one of the two distributive variables is explained independently from both the social product and the other distributive variable, and the other one is determined as a residual.

Both the classical economists and Marx considered the real wage as constituting the independent or 'given magnitude' in the relation between the two distributive variables, maintaining that its normal level is determined by 'subsistence'. Normal profits, reckoned gross of interest, are determined as a residual, on the basis of the dominant techniques of production. Given the dominant techniques, the level of the wage rate is thus the only circumstance upon which the normal rate of profit depends and no increase in the latter can be conceived of but through a fall in the former.

Wages and Profits

It is in the context of this relation between wages and profits that the problem of value arises within the classical theory. All the surplus product of the annual labour of the economy, exceeding the portion absorbed by labour itself in the form of wages, must be divided among the individual capitalists according to the capitals they have employed in production. It is the very task of relative prices ('natural prices' or 'prices or production') to ensure such proportional division of the profit share of the surplus, and in order to perform their task relative prices are bound to change in the face of any increase or fall in the quantities of the various commodities accruing to the labourers as wages. This change in relative prices, and in the value of the social product, which must necessarily take place whenever nothing changes but distribution, makes it difficult to determine the effect on profits of a rise and fall in wages; it obscures the inverse relationship between wages and profits which would be apparent if output and its means of production were the same in kind, or if their values remained unaffected by changes in the division of the product. Hence Ricardo's search for a measure of value which would be invariant to changes in wages (Ricardo 1821, ch. I, sections IV, V and VI; Sraffa 1951, pp. xlviii–xlix); hence also, Marx's determination of the general rate of profit *before and independently from* the 'prices of production', on the basis of magnitudes (the quantities of labour bestowed in the production of the relevant heterogeneous aggregates of commodities) invariant to changes in the division of the product (Marx 1894, ch. 9).

Only recently was a solution provided (Sraffa 1960) to the difficulties inherent in the theory of value that were left unresolved by Ricardo and Marx. The picture outlined above, however, points to a clear subordination of the problem of value to the determination of distribution. This contrasts sharply with the dominant supply-and-demand approach, where the theory of value – the conception of equilibrium prices as allocators of given factor endowments and their determination simultaneously with normal outputs and the equilibrium prices of factor services (distribution) – comes almost to coincide with economics itself.

As mentioned above, the real wage rate is explained by the classical authors in terms of 'subsistence'. They included in this notion 'not only the commodities which are indispensably necessary for the support of life, but whatever the custom of the country renders it indecent for creditable people, even of the lowest order, to be without', and 'the want of which would be supposed to denote that disgraceful degree of poverty, which no body can well fall into without extreme bad conduct' (Smith 1776, vol. 2, p. 399). Their conception, in other words, was that the normal wage rate 'depends not merely upon the physical, but also upon the historically developed social needs, which become second nature. But in every country, at a given time, this regulating average wage is a given magnitude' (Marx 1894, p. 859; cf. also Torrens 1815, pp. 62–3).

The classical authors also ascribed to the conditions of competition on the labour market the possibility of influencing real wages for fairly long periods of time, and hence of causing shifts away from the normal distribution of income between capitalists and workers. Smith referred to the possibility that under certain circumstances, connected with the pace of accumulation and the growth in productivity of labour, 'the scarcity of hands' or a 'scarcity of employment' may move the wage above or below the normal average level (Smith 1776, vol. 1, pp. 77 and 80). Starting from Smith's analysis, Marx went on to consider the movements of wages in the periodic alternations of the industrial cycle as regulated ' by the varying proportions in which the working-class is divided into active and reserve army, ... , by the extent to which it is now absorbed, now set free' (Marx 1883, p. 596).

Normal wages having been explained in terms of subsistence, the normal rate of profit must be determined as a residual on the basis of the dominant techniques of production. Those firms which, within each sphere of production, employ more backward or more advanced techniques than the dominant ones, earn profits that are respectively smaller or greater than normal.

In this conception, the conditions of competition amongst capitalists do not have any role to play as regulator of the normal distribution of income between wages and profits. It is easy to see on the basis of Sraffa's price equations (Sraffa 1960, paras 1–4) that, given the wage in terms of specified necessaries and the methods of production, if there is a surplus product in the economy then the system necessarily determines, together with prices, also a positive general rate of profit which no competition whatsoever among capitalists can eliminate or change. If real wages, in other words, determined by historical and social conditions independently from prices and from the rate of profit, absorb only a part of the net product of the economy, it is simply impossible for competition, however intense it may be, to determine prices such as to render nil or 'as low as possible' what remains of the value of the product after the means of production have been reintegrated and the wages paid.

It is true that the competition amongst the owners of capital plays an important role Smith's theory: he makes the level of the 'natural' rate of profit depend on it. But this is precisely where the basic contradiction in Smith's theory may be seen. On the one hand he considers the real wage to be determined by subsistence; on the other he maintains that the rate of profit is determined by competition amongst capitalists, which, by growing more intense as accumulation proceeds, would make 'the ordinary rate of profit as low as possible' (Smith 1776, vol. 1, p. 106). In short, his reasoning proceeds as if *both* distributive variables could be determined independently from each other.

Leaving aside Smith's contradiction, it can be affirmed that in classical and Marxian theory competition is envisaged essentially as the mechanism whereby, in each sphere of production, a single price tends to be established: the price that enables the means of production to be reintegrated on the basis of the dominant production techniques, and wages and profits to be paid at their normal rates. These latter must be explained independently from competition, and, as Marx puts it, it is they that regulate competition, rather than being regulated by it (Marx 1894, p. 865). The competition amongst firms within each sphere of production and the free transferability of capital from one sphere to another – hence the process whereby profit rates gravitate towards their respective normal levels – may be impeded by the presence of monopoly elements in this or that sphere of production. This however will affect the division of profits amongst the particular stocks making up social capital, but not the normal distribution of net output between wages and profits (Marx 1894, p. 861).

Interest and Profits

Profits on capital employed in production normally include, according to the classical economists, besides interest, also a remuneration for the 'risk and trouble' of productively employing it, or what may be termed a normal profit of enterprise. Production and accumulation would not continue,

Ricardo argues, if the profits of the farmers and the manufacturers were 'so low as not to afford an adequate compensation for their trouble and the risk which they must necessarily encounter in employing their capital productively' (Ricardo 1821, p. 122). Such 'adequate compensation' will be different in the various employments of capital, according to 'any real or fancied advantage which one employment may posses over another' (Ricardo 1821, p. 90). On the basis of this conception, natural prices will have to be such as to ensure that, in each sphere of production, what remains of the value of the product after deducting wages and the replacement of the means of production, is sufficient to 'adequately' remunerate the 'risk and trouble' and pay interest at an uniform rate. It can thus be said that interest and profit of enterprise are conceived in the classical analysis as the two magnitudes into which normal profits – determined by real wages and production techniques – resolve themselves.

The money rate of interest emerges from this picture as a magnitude subordinate to the normal rate of profit, being ultimately determined by those real forces, the real wage rate and production techniques, which explain the course of the normal rate of profit. But what if actual experience did not validate the conception of the money rate of interest as a subordinate phenomenon? A few significant modifications would be called for within the classical–Marxian approach to distribution, if it had to be acknowledged that the level of the rate of interest in any one country is strongly influenced by circumstances which have nothing to do with the real forces regarded by the classical economists as governing the rate of profit. These modifications, as will be apparent from the determination of distribution outlined below, would lead to a view of the real wage as the residual rather than the independent or 'given' variable in the relation between profits and wages.

It is important to notice that the replacement of the wage by the rate of profit as the independent distributive variable is fully compatible with the surplus approach to distribution (cf. Garegnani 1984, pp. 320–2). The concept of profits as surplus product is not under discussion when asking which of the two distributive variables should be regarded as 'given' in the present reality of the capitalist economy. The question is whether the relations that workers and capitalists establish with one another tend *primarily* to act upon the real wage or upon the rate of profit, once the view is abandoned that real wages consist of the necessary subsistence of the workers and the possibility of variations in the division of the social surplus is admitted.

Actual experience seems in fact to validate the conception of an autonomous determination of the money rate of interest – autonomous in the sense that interest rates *do* experience lasting changes which are very reasonably explainable without any need to refer to a *primum movens* represented by changes in the normal profit rate. Interest rates in any one country depend directly on monetary policy; interest rate policy decisions, however, are taken under a wide range of constraints having different weights both amongst the various countries and for the same country at different times: external constraints, monetary and fiscal constraints, distributive constraints. The important point is that interest rate policies, both in the short and in the long run, do not appear to be constrained by a predetermined normal profitability of capital. Once this point is acknowledged, then, given the necessary (and generally admitted) long-run connection between the rate of interest and the rate of profit, it will also be acknowledged that it is the former which 'sets the pace' and that the latter will have to adapt itself. On this basis, one can proceed to discover the actual mechanism whereby the causation occurs and to study its implications (see Pivetti 1985).

The actual mechanism whereby lasting changes in interest rates are susceptible of causing corresponding changes in normal profit rates, can be understood by following a three-stage line of reasoning. The first stage simply consists in regarding competition as the mechanism by which prices tend to be equated to normal costs. The second stage of the reasoning consists in looking at the rate of interest as a determinant of production costs, together with money wages and production techniques. Thus, lasting changes in interest rates *constitute* changes in normal costs,

which, *ceteris paribus*, will result in corresponding changes of the price level. The third stage of the reasoning comes about as a consequence of the first two: by the competition amongst firms within each industry, a lasting change in interest rates causes a change in the same direction in the level of prices in relation to the level of money wages, thereby generating changes in income distribution.

The rate of interest thus emerges from our picture as the regulator of the ratio of prices to money wages. The reader will note the main difference between this view and the so-called post-Keynesian theory of distribution: whilst in that theory changes in the level of prices in relation to the level of money wages are determined by changes in aggregate demand, according to the present explanation of distribution they are determined by lasting changes in interest rates.

By taking into consideration also the excess of profit over interest, or profit of enterprise, our conception of the rate of interest as the regulator of the ratio of prices to money wages requires us to assume that lasting changes in the rate of interest do not tend, and are not likely, to be associated with opposite changes in the normal profit of enterprise. This assumption is largely consistent with classical conceptions as regards the normal excess of profit over interest: if profit *does* normally exceed interest (if competition, that is, does not tend to equalize profit and interest), then the excess of the former over the latter must cover objective elements of 'risk and trouble' or elements which are regarded as objective by the majority of the investing public. By taking into account all such elements, we can say that the normal rate of profit in each particular production sphere will be arrived at by adding up *two* autonomous components: the long-term rate of interest or 'pure' remuneration of capital, plus the normal profit of enterprise or the remuneration for the 'risk and trouble' of productively employing capital in that sphere of production. Provided this remuneration is a sufficiently stable magnitude, lasting changes in the rate of interest will cause corresponding changes in profit rates, and inverse changes in the real wage.

Real Wages as a Residue

As we saw above, interest and profit of enterprise are conceived by the classical economists as the two magnitudes into which normal profits resolve themselves, whereas, according to our view, the same two magnitudes should rather be regarded as the *determinants* of the rate of profit. Given the money wage, the real wage appears here as a residue on the basis of the price level reflecting the dominant techniques in the different spheres of production and the normal profit rate determined in each sphere in the way we have just indicated. From this determination of distribution, quite different views from the classical ones may be developed concerning the role of competition amongst capitalists.

Since in our view the real wage constitutes the residual variable, the presence of monopoly elements in this or that sphere of production may affect not only the division of profits amongst the different employments of capital, but also the distribution between profits and wages. Given in fact the money wage, the possibility for some commodities to obtain a monopoly price which rises above the 'price of production' will translate into a ratio price-level/money wage which will be higher than it would be if there were no monopoly elements, and hence into a lower real wage. Assuming the long-term rate of interest to be unaffected by the presence of monopoly elements, it follows that lasting effects of the conditions of competition on distribution may only be obtained in one direction: higher profits than normal. For the long-term interest rate and the normal remuneration of 'risk and trouble' establish, in each sphere of production, the minimum or necessary level below which the profit rate cannot go, over the long run, however intense one may suppose the forces of competition to be.

The possibility must also be admitted that the conditions of competition influence the normal profit rate via the long-term interest rate. At the root of this possible influence of competition there is the fact that the level of the real wage constitutes *in any case* an important constraint on the freedom of monetary policy to establish the level of interest rates. To acknowledge that lasting variations in

the rate of interest determine variations in the normal distribution between profits and wages is not to concede that the real wage may move to any level whatsoever. In each concrete situation, it would be hard to carry on the productive process in an orderly manner if the real wage were lower than certain levels reflecting institutional and historical as well as economic circumstances. Thus, if the conditions of competition have a negative effect on wages – via the levels of profits of enterprise or the methods of production adopted – then beyond certain limits, which will vary from one situation to another, a compensatory effect will have to be sought in the level of interest rates.

According to our view, then, the money rate of interest should be looked on as the magnitude on which the respective powers of capitalists and workers discharge themselves *in the first place.* Wage bargaining and monetary policy are regarded as the main channels through which class relations act in determining distribution, and those relations are seen as tending to primarily act upon the profit rate, via the monetary rate of interest, rather than upon the real wage rate as maintained by both the classical economists and Marx. The level of the real wage prevailing in any given situation is the *final* result of the whole process by which distribution of income between workers and capitalists is actually arrived at.

It seems to us that in the conditions of modern capitalism it is difficult to conceive of the real wage rate as the independent or given variable in the relationship between wages and profits – the difficulty, as we see it, arising from the fact that the direct outcome of wage bargaining is a certain level of the money wage, while the price level cannot be determined before and independently from money wages. Given distribution between profits and wages, and given the methods of production, the level of prices simply depends on the level of money wages. Thus, in our picture, the long-term rate of interest enters into the determination of the price level because it contributes to regulating the ratio of the latter to the money wage – that is, distribution between profits and wages.

If instead the real wage is taken as given, the ratio of prices to money wages will be determined by the condition that it must be such as to ensure the given level of the real wage; and on this basis wage bargaining, in determining money wages, can be thought of as determining also the price level. In such a picture monetary policy plays a purely passive role – the level of the rate of interest having to accommodate to lasting changes in the ratio of prices to money wages, rather than governing that ratio. Now what we are ultimately facing here is a conception of the ratio of prices to money wages as being determined by a magnitude, the real wage rate, which is not actually known before that ratio is known. This explains in our opinion why of the two alternative propositions – that the ratio of prices to money wages depends on the real wage rate, or that the real wage rate depends on the ratio of prices to the money wage – the latter is easier to digest: in actual fact, there are no circumstances determining real wages as distinct from those acting through money wages, the level of prices and the ratio of prices to money wages.

See Also

► Surplus

Bibliography

Cairnes, J.E. 1874. *Some leading principles of political economy newly expounded*. New York: Harper & Brothers.

Cannan, E. 1917. *A history of the theories of production and distribution in English political economy from 1776 to 1848*. 3rd ed. London: King.

Garegnani, P. 1984. Value and distribution in the classical economists and Marx. *Oxford Economic Papers* 36: 291–325.

Malthus, T.R. 1827. *Definitions in political economy, preceded by an inquiry into the rules that ought to guide political economists in the definition and use of their terms; with remarks on the deviations from these rules in their writings.* Reprinted. New York: Kelley, 1971.

Marx, K. 1862–63. *Theories of Surplus Value,* vols 1–3. Moscow: Progress Publishers, 1963–71.

Marx, K. 1873. Afterword to the 2nd German edition of *Capital.* In Marx (1883).

Marx, K. 1883. *Capital: A critique of political economy,* vol. 1. 3rd edn, 1977, reprinted. London: Lawrence & Wishart.

Marx, K. 1894. *Capital: A critique of political economy,* vol. 3, 1977. London: Lawrence & Wishart.
Mill, J. 1844. *Elements of political economy.* 3rd ed, 1965. New York: Kelley.
Mill, J.S. 1848. *Principles of political economy with some of their applications to social philosophy.* Toronto/London: University of Toronto Press/ Routledge & Kegan Paul, 1965.
Pivetti, M. 1985. On the monetary explanation of distribution. *Political Economy – Studies in the Surplus Approach* 1 (2): 73–103.
Ricardo, D. 1821. Principles of political economy and taxation. In *The works and correspondence of David Ricardo*, ed. P. Sraffa, vol. 1, 3rd edn. Cambridge: Cambridge University Press, 1951.
Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations,* ed. E. Cannan. London: Methuen, 1961.
Sraffa, P. 1951. Introduction to Ricardo's *Principles*. In *Collected works of David Ricardo*, vol. 1. Cambridge: Cambridge University Press.
Sraffa, P. 1960. *Production of commodities by means of commodities*. Cambridge: Cambridge University Press.
Torrens, R. 1815. *An essay on the external corn trade*. London: Hatchard.
Torrens, R. 1821. *An essay on the production of wealth. With an appendix in which the principles of political economy are applied to the actual circumstances of this Country.* Reprinted. New York: Kelley, 1965.

Classical Economics and Economic Growth

Gavin Cameron

Abstract

The classical economists dealt with many of the issues now addressed by modern growth theories, albeit with different theoretical tools and with different perspectives. Classical analyses of the division of labour, population growth, and the difficulties when factors are in fixed supply, continue to have modern applications. However, the models they developed ran into difficulties after the 'marginalist revolution', when it became apparent that sustained technical change, abstinence and thrift by the labouring classes, and factor substitution might forestall the arrival of the stationary state.

Keywords

Balanced growth; Capital accumulation; Classical economics; Classical economics and economic growth; Diminishing returns; Distribution theory; Division of labour; Economic growth; Falling rate of profit; Fertility; Human capital; Industrial Revolution; Labour supply; Luxury; Malthus, T. R.; Marginalist revolution; Marx, K.H.; Mill, J. S.; Mortality; Population growth; Productive and unproductive labour; Ricardo, D.; Say, J.-B.; Smith, A.; Stationary state; Technical change; Value theory

JEL Classification

B1; O4; Q3

The analysis of economic growth was an important feature of the writings of the great classical economists, including Adam Smith, Thomas Malthus, David Ricardo, John Stuart Mill and Karl Marx.

To place them in their historical context is straightforward if economic history is simplified into three distinct epochs. In the first, which spanned most of human history and still obtains in some unfortunate regions, Malthusian conditions prevailed: living standards were static even though there was some population growth. In the second, which began in the middle of the eighteenth century in England, living standards showed some upward tendency and there was a demographic change as fertility rates rose and mortality rates fell, resulting in a substantial rise in population. In the third epoch, characteristic of England from the 1820s perhaps, the move to sustained economic growth provoked a shift from quantity to quality in child-rearing, and all the appurtenances of modern growth began to appear, such as human capital, professional R&D, and technical innovation.

There is much scope for discussion about what factors triggered, propagated, and enhanced such changes, and about when such changes began and whether they were smooth or discrete. For example, Mokyr (2005) argues that living standards in England rose gently between the 17th and

18th centuries due to the spread of world trade, commercialism and the rise of institutions less hostile to consumers and the industrious – nicely, this is sometimes called 'Smithian' growth. Somewhat in contrast, Allen (2001) argues that real wages did not rise significantly over that period in England, but that, since they were falling across most of Europe, the real question is what would have happened in the absence of the Industrial Revolution.

Mokyr also points out that many of the inventions associated with the eighteenth century Industrial Revolution were developed in northwest Europe, but successfully applied in England. It is not surprising that the classical economists were fascinated. Adam Smith was born in 1723, within the Malthusian growth regime, whereas Ricardo, Malthus and Jean-Baptiste Say were well placed to observe the demographic change in England and the beginnings of industry, even though England was still predominantly a rural society in the early nineteenth century. Unsurprisingly, Mill and Marx found it increasingly hard to defend Ricardian doctrines as the modern growth regime began to emerge across Europe and its offshoots in the middle of the nineteenth century.

Being products of the Enlightenment, the classical economists shared a concern for human progress that would do credit to a modern policymaker. One purpose of their analysis was to identify the forces in society that promoted or hindered progress and to provide a basis for policy and action in a time of considerable political innovation in England (including land enclosures, franchise reform, tariff reform, and the abolition of the slave trade) and revolution abroad (including land reform, the continental system, and the tumbrils). This background motivated Ricardo's campaign against the Corn Laws, as it did Malthus's concern with population growth, Smith's attacks on mercantilism, and Marx's analyses of social class.

The classical economists' work was grounded in the economic conditions of their times, and not in the abstract mathematical reasoning that appeared in economics during the marginalist revolution of the 1870s and after, popularized by Ysidro Edgeworth, William Stanley Jevons and Alfred Marshall. In contrast to more recent economic thought, the classical economists saw discussions of economic growth as being inextricably linked with discussions of the theory of value and the theory of distribution. Since their concerns were largely those of educated gentlemen of those times, they wanted to be able simultaneously to explain trade cycles, inflation and other short-run phenomena, as well as real wages and population growth and other long-run phenomena. While it is easy to see the current gap between short-run and long-run macroeconomic models as a lacuna (for example, see Solow 2005), the classical economists tended to run into problems when treating both at the same time.

> The characteristic features of what is commonly meant by industrial progress, resolve themselves mainly into three, increase of capital, increase of population, and improvements in production; understanding the last expression in its widest sense, to include the process of procuring commodities from a distance, as well as producing them. (Mill 1848, Book IV, ch. 3)

The classical economists also worried about the consumption of luxuries and the distinction between productive and unproductive labour. As Brewer (1997) discusses, this is particularly true of Adam Smith, who displays a good deal of ambivalence about luxuries:

> That portion of his revenue that a rich man annually spends is in most cases consumed by idle guests and menial servants, who leave nothing behind them in return for their consumption. That portion which he annually saves, for the sake of the profit it is immediately employed as a capital, is consumed in the same manner, and nearly the same time too, but by a different set of people, by labourers, manufacturers and artificers, who reproduce with a profit the value of their annual consumption. (Smith 1776, Book II, ch. 3)

Smith's view contrasts somewhat with that of his predecessor David Hume, whose mild approval of luxuries was based on the notion that they might encourage economic and political development. Although such notions still figure in modern debates (Greenhalgh 2005), this preoccupation with luxuries and unproductive labour turns out to be not very useful for modelling purposes, unless it is simply be taken to mean that different economic groups have different

propensities to save, which is a truism. However, even if the classical economists did not always approve of certain kinds of consumption, Smith's contention that consumption is the sole end and purpose of all production was a vast improvement on the mercantilist doctrine.

Clearly, the classical economists cannot be written off as growth theorists manqué. The technical core of modern growth theory rests upon technical change, specialization, factor substitution, and factor accumulation, with various recent theorists emphasizing the effects on these of trade, institutions, inequality, political economy, geography and population size and growth. All these issues were concerns of the classical economists, even if they used a different vocabulary.

Nonetheless, it would be fair to say that the classical economists have had only a limited direct impact on recent growth theorists. Adam Smith receives seven references in the current two-volume *Handbook of Economic Growth* (Aghion and Durlauf 2005). Malthus a very respectable 13, while of the other classical economists only Ricardo merits a single mention. Interestingly, an even older economist, William Petty from the seventeenth century, is often quoted in writing about the effect of population size on inventiveness in the scale effect literature (see Jones 2005).

The Stationary State

The classical economists saw all around them the effects of the development of the capitalist system, most importantly, of course, the accumulation of capital, but also the introduction of new techniques. Smith analysed in great detail the process of the division of labour, but more generally the classical economists did not attempt to deal with the relationship between capital accumulation and technical change (although Marx did highlight the issue). In addition to these basic forces of economic growth, they were also interested in the increase in the supply of labour through population growth. In the case of Thomas Malthus, this interest was quite morbid.

> The power of population is so superior to the power in the earth to produce subsistence for man that premature death must in some shape or other visit the human race. (Malthus 1798)

The classical economists' analysis of the process by which capital, technology and labour grow over time led them to a common conclusion, motivated by different causes – that the process of economic growth was gradually self-attenuating and ended in a state of stagnation (the 'stationary state'):

> When the stocks of many merchants are turned into the same trade, their mutual competition naturally tends to lower its profit; and when there is a like increase of stock in all the different trades carried on in the same society; the same competition must produce the same effect in them all. (Smith 1776, Book I, ch. 4)

The principal way in which Smith envisaged a stationary state as obtaining was that the rate of profit would fall as capital accumulated in the long run due to increased competition. Smith associated this stationary state with the position of China, which he described as being one of the most fertile and industrious countries, but also as having low wages and having been long stationary. There is tension in the Wealth of Nations between three separate points: first, his worries about the falling rate of profit; second, his worries that wages could fall to a subsistence level; and third, his description of net saving creating higher levels of output. This shows that although the economic system he describes is very complex, it tends to neglect both the feedback between profits and saving, and substitution between capital and labour.

Some controversy exists about the origin of the idea of 'diminishing returns', although it certainly appears in the writings of Jacques Turgot in the eighteenth century. The early nineteenth-century English economists certainly saw the idea in action with the expansion of cultivated land in England during the Napoleonic Wars. Subsequently, the idea comes to life in Ricardo's 'corn' model. Modern presentations of this model are plentiful (see for example, Kaldor 1956; Pasinetti 1960; Samuelson 1978;

discussions in Glyn 2004). The presentation here follows Bhaduri and Harris (1987).

Suppose that there is a single product, 'corn', produced in a capitalist agricultural economy. Land differs in its fertility and labour is applied in fixed proportions to land of diminishing fertility. The supply of labour is perfectly elastic at some fixed real wage equal to 'subsistence' (this is clearly an extreme form of the Malthusian hypothesis; see for example, in Samuelson 1978, and discussion in Brezis and Young 2003). Total output is distributed between rent paid to landlords, profits to capitalists, and wages. The level of land rent can then be shown to be determined by the difference between the average and marginal product of labour at the prevailing level of employment, and profits are the residual after rent and wages are paid (equal to the marginal product of labour minus the wage, times employment). Although there is a variety of Ricardian schemes for the determination of saving (and hence capital accumulation in a closed economy with no consumption loans), a typical presentation takes saving to be a constant proportion of profits, so the rate of accumulation is uniquely dependent upon the profit rate.

However, as employment growth proceeds, the marginal product of labour falls and so must the profit rate. The system asymptotically approaches a stationary state when the profit rate is so low that accumulation ceases (the 'minimum acceptable rate of profit'). What happens is that capitalists find themselves squeezed between the diminishing product of labour and the need to pay the going wage rate, and paying out an increasing share of output as rent to landlords. There is thus a conflict between landlords and capitalists.

In the absence of technical change, the possibility that landlords or workers could themselves become savers, or substitution away from that resource, any other fixed resource would play the same role. Samuelson (1978) notes that neither Ricardo nor Marx was so naive as to believe literally in fixed proportions between capital goods and labour, but their models were unable fully to reflect this complexity.

Mill provides both a summary and a synthesis of previous writers, drawing particularly on Ricardo:

> On the whole, therefore, we may assume that in a country such as England, if the present annual amount of savings were to continue, without any of the counteracting circumstances which now keep in check the natural influences of those savings in reducing profit, the rate of profit would speedily attain the minimum, and all further accumulation of capital would for the present cease. (Mill 1848, Book IV, ch. 4)

Mill contradicts Smith's assertion that competition is the cause of the falling profit rate and proposes instead a form of diminishing returns to capital, provided by limits to the 'field of employment' of capital. He then explicitly links capital accumulation with saving and notes that there is some minimum rate of profit, below which capital accumulation cannot take place. However, he does propose four mechanisms by which the stationary state may be overcome: first, that capital may be wasted during speculative booms; second, through improvements in production; third, through an expansion of foreign trade, and fourth, through the export of capital to other countries.

The second is the one that resonates with modern growth theory, although Mill muddies the waters with a contradictory passage about why an improvement in the production of luxuries (such as lace and velvet) will affect capital accumulation through a different mechanism.

Marx was also a firm believer in this movement towards a stationary state, exemplified by what he called the falling tendency of the rate of profit (FTRP). In the Marxian scheme, the FTRP is one of the main sources of crises under capitalism. Writers in this tradition usually understate the ability of technical progress to reliably prevent such crises and overstate the role of the business cycle in long-run development. Not every slump or financial crash heralds the end of capitalism. But on the former point, Marx was writing at an early stage of the sustained growth era, largely before the existence of large-scale industrial processes and certainly before professional R&D laboratories (see Glyn 2006, for a discussion of whether the entry of China and India into the global economy might

presage a return to a Marxian era of growth). In such an era, technical innovation may well have appeared more uncertain and less widespread than it would later appear, or, to use Harberger's analogy, more like mushrooms popping up here and there than like yeast leavening the entire economic process (Harley 2003).

It can be seen that the classical economists were much more concerned about the stationary state than if it just represented an equilibrating tendency in a long-run growth model à la Solow where capital deepening slows in the absence of technical change (this is clear from Sweezy 1942, ch. 9). Nonetheless, in the idea of the stationary state (and from Mill's view that he was considering the 'dynamics' of the economy, having dealt with the 'statics'), it is possible to see the seed-corn of the Solow model, once economists such as Marshall, Frank Ramsey, Charles Cobb, and Paul Douglas had laid further foundations.

In contrast, classical theories of growth qua theories of growth became increasingly marginal as the nineteenth century wore on (although of course, Marxian and Marxist analysis remained influential for much longer). The Swedish unemployment of the early 1920s prompted Knut Wicksell to write three articles from a neo- Malthusian standpoint, one of which, entitled 'Ricardo on Machinery and the Present Unemployment', he submitted to the *Economic Journal*. John Maynard Keynes, the editor of the journal, rejected the paper, arguing 'that any treatment of this topic at the present day ought to bring in various modern conceptions for handling the problem and the time has gone by for a criticism of Ricardo on purely Ricardian lines' (J.M. Keynes, quoted in Jonung 1981). In the end, even Piero Sraffa's remarkable work, *Production of Commodities by Means of Commodities* (1960), was not enough to revive Ricardian analysis, although some still see neoclassical economics as its direct descendant (Hollander 1995).

Conclusion

Classical economists are often regarded as 'pessimistic' in their forecasts of the future development of the economy, and came in for heavy criticism from the unlikeliest of sources, the Romantic poets and literary critics such as Ruskin. This kind of trahison des clercs of poets and authors against a changing social order and increasing commercialization is familiar to a modern reader of tracts against global capitalism, and equally well grounded in theory and evidence.

The classical economists' search for a 'theory of value' and a 'theory of distribution' was an attempt to understand the significant economic, political, and social changes of their times, as well as an attempt to understand what would happen in the long run in those economies. There is much to be learnt from their analyses, both as an indicator of the conditions of the times (that is, the importance of land as a factor of production) and also as a precursor to the future development of the theory of economic growth. Without the analytical apparatus that arose during the marginalist revolution (such as production functions and utility functions), their analyses were hampered, but a number of the features that drive modern models of growth made their first appearance in the writings of the classical economists. For example, the importance of the division of labour, technical progress and the role of population growth, as well as the idea of diminishing returns, all feature prominently in modern models.

What is lacking from the classical accounts is the notion of a balanced growth path. The classical economists largely concluded that, in the long run, economies would tend towards a stationary, stagnant state. They emphasized the ability of population growth to keep wages at subsistence level, the notion that capital could only be accumulated out of profits, and the central role of land as a factor of production. In this sense, their analytical scheme is flawed. Economic progress has shown that the possibility of investment in human capital can lead to a demographic shift whereby households choose 'quality' over 'quantity' in their reproductive choices; that saving by workers can be an important source of capital accumulation; and that factor substitution tends to prevent the inexorable rise in the price of any factor, even if it is in fixed supply.

See Also

- ▶ Balanced Growth
- ▶ Malthus, Thomas Robert (1766–1834)
- ▶ Marx, Karl Heinrich (1818–1883)
- ▶ Mill, John Stuart (1806–1873)
- ▶ Ricardo, David (1772–1823)
- ▶ Smith, Adam (1723–1790)

Acknowledgment The author would like to thank Robert Allen, Julia Cartwright, Mary Dixon-Woods, Marcel Fafchamps, Nicholas Fawcett, Andrew Glyn, Mark Koyama, Silvia Palano, and Jonathan Temple for helpful comments on a preliminary draft.

Bibliography

Aghion, P., and S. Durlauf. 2005. *Handbook of economic growth*. Amsterdam: North-Holland.

Allen, R. 2001. The great divergence in European wages and prices from the Middle Ages to the First World War. *Explorations in Economic History* 38: 411–447.

Bhaduri, A., and D. Harris. 1987. The complex dynamics of the simple Ricardian system. *Quarterly Journal of Economics* 102: 893–902.

Brewer, A. 1997. Luxury and economic development: David Hume and Adam Smith. *Scottish Journal of Political Economy* 45: 78–98.

Brezis, E., and W. Young. 2003. The new views on demographic transition: A reassessment of Malthus's and Marx's approach to population. *European Journal of the History of Economic Thought* 10: 25–45.

Glyn, A. 2004. *The corn model, gluts and surplus value*. Working Paper No. 194, Department of Economics, Oxford University.

Glyn, A. 2006. Will Marx be proved right? *Oxonomics* 1: 13–16.

Greenhalgh, C. 2005. Why does market capitalism fail to deliver a sustainable environment and greater equality of incomes? *Cambridge Journal of Economics* 29: 1091–1109.

Harley, K. 2003. Growth theory and industrial revolutions in Britain and America. *Canadian Journal of Economics* 36: 809–831.

Hollander, S. 1995. *Collected essays I: Ricardo. The 'new view'*. London: Routledge.

Jones, C. 2005. Growth and ideas. In Aghion and Durlauf (2005).

Jonung, L. 1981. Ricardo on machinery and the present unemployment: An unpublished manuscript by Knut Wicksell. *Economic Journal* 91: 195–198.

Kaldor, N. 1956. Alternatives theories of distribution. *Review of Economic Studies* 28: 83–100.

Malthus, T.R. 1798/1999. *An essay on the principle of population*. Oxford: Oxford World's Classics.

Mill, J.S. 1848/1985. *Principles of political economy, with some of their applications to social philosophy*. London: Penguin Classics.

Mokyr, J. 2005. Long-term economic growth and the history of technology. In Aghion and Durlauf (2005).

Pasinetti, L. 1960. A mathematical formulation of the Ricardian system. *Review of Economic Studies* 27: 78–98.

Samuelson, P. 1978. The canonical classical model of political economy. *Journal of Economic Literature* 16: 1415–1434.

Samuelson, P. 1988. Mathematical vindication of Ricardo on machinery. *Journal of Political Economy* 96: 274–282.

Smith, A. 1776/1976. *An inquiry into the nature and causes of the wealth of nations*, ed. R.H. Campbell and A.S. Skinner, 2 vols. Oxford: Oxford University Press.

Solow, R. 2005. Introduction: Growth in retrospect and prospect. In Aghion and Durlauf (2005).

Sraffa, P. 1960. *Production of commodities by means of commodities: Prelude to a critique of economic theory*. New York: Cambridge University Press.

Stigler, G. 1958. Ricardo and the 93% labor theory of value. *American Economic Review* 48: 357–367.

Sweezy, P. 1942. *The theory of capitalist development*. New York: Monthly Review Press.

Classical Growth Model

Donald J. Harris

Abstract

The classical economists provided an account of the broad forces that influence economic growth and of the mechanisms underlying the growth process, stressing accumulation and productive investment of a part of the social surplus in the form of profits. Changes in the rate of profit were decisive for analysis of the long-term evolution of the economy. The analysis indicated that in a closed economy there is an inevitable tendency for the rate of profit to fall. In this article, the essential features of the classical analysis of the accumulation process are presented and formalized in terms of a simple model.

Keywords

Capital accumulation; Capitalism; Class; Classical economics; Classical growth model; Corn

as basic commodity; Corn Laws; Diminishing returns; Dismal science; Distribution theories, classical; Division of labour; Economic growth; Falling tendency of the rate of profit; Household production; Labour productivity; Labour supply; Malthus, T. R.; Malthus's theory of population; Marx, K. H.; Physiocracy; Population growth; Progress; Rate of profit; Reserve army of labour; Ricardo, D.; Smith, A.; Stationary state; Subsistence; Surplus; Technical change; Value; Wages fund

JEL Classifications
O4

Analysis of the process of economic growth was a central feature of the work of the English classical economists, as represented chiefly by Adam Smith, Thomas Malthus and David Ricardo. Despite the speculations of others before them, they must be regarded as the main precursors of modern growth theory. The ideas of this school reached their highest level of development in the works of Ricardo.

The interest of these economists in problems of economic growth was rooted in the concrete conditions of their time. Specifically, they were confronted with the facts of economic and social changes taking place in contemporary British society as well as in previous historical periods. Living in the 18th and 19th centuries, on the eve or in the full throes of the Industrial Revolution, they could hardly help but be impressed by such changes. They undertook their investigations against the background of the emergence of what was to be regarded as a new economic system – the system of industrial capitalism. Political economy represented a conscious effort on their part to develop a scientific explanation of the forces governing the operation of the economic system, of the actual processes involved in the observed changes that were going on, and of the long-run tendencies and outcomes to which they were leading.

The interest of the classical economists in economic growth derived also from a philosophical concern with the possibilities of 'progress' an essential condition of which was seen to be the development of the material basis of society. Accordingly, it was felt that the purpose of analysis was to identify the forces in society that promoted or hindered this development, and hence progress, and consequently to provide a basis for policy and action to influence those forces. Ricardo's campaign against the Corn Laws must obviously be seen in this light, as also Malthus's concern with the problem of population growth and Smith's attacks against the monopoly privileges associated with mercantilism.

Of course, for these economists, Smith especially, progress was seen from the point of view of the growth of national wealth. Hence, the principle of national advantage was regarded as an essential criterion of economic policy. Progress was conceived also within the framework of a need to preserve private property and hence the interests of the property-owning class. From this perspective, they endeavoured to show that the exercise of individual initiative under freely competitive conditions to promote individual ends would produce results beneficial to society as a whole. Conflicting economic interests of different groups could be reconciled by the operation of competitive market forces and by the limited activity of 'responsible' government.

As a result of their work in economic analysis the classical economists were able to provide an account of the broad forces that influence economic growth and of the mechanisms underlying the growth process. An important achievement was their recognition that the accumulation and productive investment of a part of the social product is the main driving force behind economic growth and that, under capitalism, this takes the form mainly of the reinvestment of profits. Armed with this recognition, their critique of feudal society was based on the observation, among others, that a large part of the social product was not so invested but was consumed unproductively.

The explanation of the forces underlying the accumulation process was seen as the heart of the problem of economic growth. Associated with accumulation is technical change as expressed in the division of labour and changes in methods of production. Smith, in particular, placed heavy

emphasis on the process of extension of division of labour, but there is, in general, no systematic treatment of the relation between capital accumulation and technical change in the work of the classical economists. It later becomes a pivotal theme in the work of Marx and is subjected there to detailed analysis (see, for instance, Marx 1867, part 4). To these basic forces in economic growth they added the increase in the supply of labour available for production through growth of population. Their analysis of the operation of these forces led them to the common view, though they quite clearly differed about the particular causes, that the process of economic growth under the conditions they identified raises obstacles in its own path and is ultimately retarded, ending in a state of stagnation – the 'stationary state'.

The conception of the stationary state as the ultimate end of the process of economic growth is often interpreted as a 'prediction' of the actual course of economic development in 19th-century England. There is no doubt that it was for a time so regarded by some, if not all, of the economists and their contemporaries, though the weight that was assigned to this particular aspect of the conception by Ricardo himself is a matter of some dispute. What is more significant, however, is that this conception served to point to a particular social group, the landlord class, who benefited from the social product without contributing either to its formation or to 'progress' and who, by their support of the Corn Laws and associated restrictions on foreign trade, acted as an obstacle to the only effective escape from the path to a stationary state, that is, through foreign trade.

In examining the work of the classical economists we find also that problems of economic growth were analysed through the application of general economic principles, viewing the economic system as a whole, rather than in terms of a separate theory of economic growth as such. These principles were such as to recognize basic patterns of interdependence in the economic system and interrelatedness of the phenomena of production, exchange, distribution and accumulation. In sum, what we find in classical economic analysis is a necessary interconnection between the analysis of value, distribution and growth. Because of these interconnections it was by no means possible to draw a sharp dividing line between the inquiry into economic growth and that into other areas of political economy. As Meek (1967, p. 187) notes:

> To Smith and Ricardo, the macroeconomic problem of the 'laws of motion' of capitalism appeared as the primary problem on the agenda, and it seemed necessary that the whole of economic analysis – including the basic theories of value and distribution – should be deliberately oriented towards its solution.

Distribution of the social product was seen to be connected in a definite way with the performance of labour in production and with the pattern of ownership of the means of production. In this regard, labour, land, and capital were distinguished as social categories corresponding to the prevailing class relationships among individuals in contemporary society: the class of labourers consisted of those who performed labour services, landlords were those who owned titles or property in land, and capitalists were those who owned property in capital consisting of the sum of exchangeable value tied up in means of production and in the 'advances' which go to maintain the labourers during the production period. Each class received income or a share in the product according to specified rules: for the owners, the rule was based on the total amount of property which they owned – so much rent per unit of land, so much profit per unit of capital (and, for the class of finance capitalists or 'rentiers' who lent money at interest, so much interest per unit of money lent). For labourers it was based on the quantity of labour services performed: so much wages per hour.

Accumulation and distribution were seen to be interconnected through the use that was made by different social classes of their share in the product. Basic to this view was a conception, taken over from the Physiocrats, of the social surplus as that part of the social product which remained after deducting the 'necessary costs' of production consisting of the means of production used up and the wage goods required to sustain the labourers employed in producing the social product. This

surplus was distributed as profits, interest and rent to the corresponding classes of property owners. For the classical economists, the possibility of accumulation was governed by the size and mode of utilization of this surplus. Accordingly, their analysis placed emphasis upon those aspects of distribution and of the associated class behaviour which had a direct connection with the disposal of the surplus and therefore with growth. In particular, it was assumed that, typically, workers consumed their wages for subsistence, capitalists reinvested their profits and landlords spent their rents on 'riotous living'. On the other side, accumulation would also influence the distribution of income as the economy expanded over time.

It was this absolutely strategic role of the size and use of the surplus, viewed from the perspective of the economy as a whole and of its process of expansion, which dictated the significance of the distribution of income for classical economic analysis. Thus, for Ricardo especially, investigation of the laws governing distribution became the focus of analysis. In a letter to Malthus, Ricardo wrote (*Works*, VIII, pp. 278–9): 'Political Economy you think is an inquiry into the nature and causes of wealth; I think it should rather be called an inquiry into the laws which determine the division of the produce of industry among the classes which occur in its formation.' What was of crucial significance in this connection was the rate of profits because of its connection with accumulation, both as the source of investment funds and as the stimulus to further investment.

Having 'got rid of rent' as the difference between the product on marginal land and that on intra-marginal units, the Ricardian analysis focused on profits as the residual component of the surplus. Under the simplifying conditions on which the analysis was constructed, there emerged a very clear and simple relationship between the wage rate and the overall rate of profits, determined within a single sector of the economy – the corn-producing sector. The special feature of corn as a commodity was that it could serve both as capital good (seed corn) in its own production and as wage good to be advanced to the workers. With the wage rate fixed in terms of corn, the rate of profit in corn production is uniquely determined as the ratio of net output of corn per man minus the wage to the sum of capital per man consisting of seed corn and the fund of corn as wage good. Competition ensures that the same rate of profit enters into the price of all other commodities that are produced with indirect labour. The overall rate of profits, determined in this way, varies inversely with the corn wage. But, as soon as it is recognized that the wage and/or the capital goods employed in corn production consist of other commodities besides corn, the rate of profits can no longer be determined in this way. For the magnitude of the wage and of the total capital then depends on the prices of those commodities, and these prices incorporate the rate of profit. Attention then has to be directed to explaining the rate of profit by taking account of the whole system of prices. For this purpose the theory of value is called upon to provide a solution and Ricardo struggled with this problem until the end of his life. An elegant solution has been worked out by Sraffa (1960) which shows that, in a system of many produced commodities, with the real wage rate given at a specified level, the rate of profit is determined by the given wage and the conditions of production of the commodities that are 'basics'. It so happens that Ricardo's case of corn is just such a 'basic' commodity in the strict sense that it enters directly and indirectly into the production of every commodity including itself.

The core idea that competition among firms under capitalist conditions tends to produce uniformity of profit rates across all markets remains problematical, especially in the dynamic real-world context of changing technology with various forms of factor immobility and barriers to entry (Harris 1988).

Given the perceived centrality of the rate of profit in a capitalist economy, for classical political economy it becomes a crucial problem in the theory of economic growth to account for movements in the rate of profit associated with the process of capital accumulation and development of the economy. Such movements are a decisive reference point for understanding the long-term evolution of the economy. The classical answer to this problem, as worked out most coherently by

Ricardo, is that in a closed economy there is an inevitable tendency for the rate of profit to fall in the course of the accumulation process and, hence, that the accumulation process itself is brought to a halt by its own logic.

Marx was later to propose this falling tendency of the rate of profit (FTRP) as a *law*. He considered it to be 'the most important law of modern political economy' (1973, p. 748; 1894, part 3). He was, of course, following in the tradition of the classical economists in which the same idea had been firmly entrenched, though supported on different grounds. But, interestingly enough, it is also the case that there exists a distinct conception of a FTRP within neoclassical theory (Harris 1978, ch. 9; 1981). In Keynes, as well, the idea is embodied in his projection of the long-term prospects for capitalism resulting in the 'euthanasia of the rentier' (1936, pp. 375–6). In Schumpeter (1934), it occurs in the form of the idea that the profitability of innovations tends inevitably to be eroded so that the economy settles back to the conditions of the 'circular flow' in the absence of new innovations. Though it is based in each case on quite different foundations, this conception is one of the most striking and persistent uniformities across different schools of economic thought. (For a discussion of the long history of the idea of a falling rate of profit, see Tucker 1960.)

A Model of Accumulation

The essential features of the classical argument regarding the accumulation process can be exhibited with a simple model adapted from Kaldor (1956) and Pasinetti (1960). This model formalizes the Ricardian conception of an agricultural economy producing a single product, 'corn', under capitalist conditions. Land is of differing fertility and labour is applied in fixed proportion to less and less fertile land. Accordingly, the average and marginal product of labour falls as the margin of cultivation is extended through capital accumulation and increase of employment on the land. The system may indifferently be assumed to expand on the extensive or intensive margins of available land. Also, it does not matter for this analysis that there exists any production outside agriculture. It would turn out, in any case, that the overall average rate of profit for the economy as a whole is determined by the agricultural rate of profit or, in the general case, by the conditions of production of 'basics' (see Sraffa 1960; Pasinetti 1977). Of course, in a system with many produced commodities, it is not possible to define 'less fertile land' independently of the rate of profit (Sraffa 1960). However, the problem does not arise in this simplified model of a corn-producing economy. We deliberately abstract from complications associated with the Malthusian population dynamics. This is perhaps the most problematic feature of the classical conception and we return to it below. Meanwhile, it is simply assumed, as in Lewis (1954), that a labour force is in perfectly elastic supply at some conventionally fixed real wage rate equal to 'subsistence'.

Let the production function relating output Y to labour input L be

$$Y = F(L) \quad \begin{array}{l} F(0) \geqq 0 \\ F' > w^* > 0 \\ F'' < 0 \end{array} \tag{1}$$

which satisfies the law of diminishing returns and allows for the existence of a surplus product above the 'subsistence' wage-rate w^*. Total capital K consists entirely of wages W (the 'wage fund') advanced at the beginning of the production period to hire labour. Thus

$$K = W = wL. \tag{2}$$

We are here, for simplicity, neglecting capital as seed-corn, and inputs of fixed capital are ignored. Total output is distributed between payment of rent R to landlords, profits P to capitalists, and replacement of the wage fund:

$$Y = R + P + W. \tag{3}$$

Given the margin of cultivation reached at any time, the level of land rent is determined as the difference between the average and marginal product of labour at the prevailing level of employment:

$$R = \left(\frac{F(L)}{L} - F'\right)L. \tag{4}$$

Profit emerges as the residual

$$P = (F' - w^*)L. \tag{5}$$

It follows that the rate of profit r is determined from

$$r = \frac{P}{W} = \frac{F'}{w^*} - 1. \tag{6}$$

It is the dynamics of the wage fund which represents the process of accumulation in this model. Accumulation of capital consists of the growth of the wage fund with a corresponding increase of employment. Additions to the wage fund come entirely from investment of capitalists' profits since the spendthrift landlords consume their share of the surplus. If the capitalists invest a proportion of profits equal to α, then

$$\Delta W = \alpha P \quad 0 < \alpha < 1. \tag{7}$$

The proportion α need not be a constant. It could vary in a manner dependent on the rate of profit as suggested by Ricardo's idea that

> [the capitalists'] motive for accumulation will diminish with every diminution of profit, and will cease altogether when their profits are so low as not to afford them an adequate compensation for their trouble and the risk which they must necessarily encounter in employing their capital productively. (*Works*, I, p. 122)

In that case we have

$$\alpha = \alpha(r) \quad \alpha' > 0 \alpha(r^*) = 0 \tag{8}$$

where r^* is the capitalists' minimum acceptable rate of profit. By definition the rate of capital accumulation is $g = \Delta W/W$, and from (6), (7), and (8) it follows that

$$g = \alpha(r) \cdot r. \tag{9}$$

Thus, the rate of accumulation is uniquely dependent on the profit rate.

The movement in the profit rate as accumulation proceeds can be derived from (6). Evidently, as employment increases the marginal product of labour falls. The rate of profit must therefore fall. It continues to fall as long as there is any increment to the wage fund so as to employ extra labour on the available land. The process comes to a halt when the profit rate is so low that accumulation ceases. The economy is then at the stationary state.

In this model, the capitalists are caught between, on the one hand, the diminishing productivity of labour as the margin of cultivation is extended and, on the other, the need to pay the ongoing wage rate in order to secure labour for employment. As the productivity of labour falls on the marginal land the pressure of land rent increases for the existing intra-marginal units. The capitalists must therefore pay out an increasing share of the surplus to the landlords. In this way they gradually lose command over the investible surplus of the economy to the landlord class. This distributional conflict between the landlord class and the capitalists constitutes a central feature of the process that drives the economy towards its ultimate stationarity. The impenetrable barrier in the process is the diminishing fertility of the soil. More generally, it is the limitation of natural resources, in this case land, which brings the process to a halt. In this respect the classical model is a particular case of resource-limited growth. Any other limited resource would have the same effect, through increasing 'rents' for that resource. At the same time, this consequence is also the product of the capitalists' own actions in relentlessly seeking to expand the size of their capital.

The underlying dynamic process which expresses this conflictive evolution of capitalist accumulation has usually been assumed in the literature to converge towards the stationary state (see Pasinetti 1960; Samuelson 1978). Some reservation on this question of convergence was originally expressed by Hicks and Hollander (1977) and followed up by Gordon (1983). Subsequent discussion by Casarosa (1978), Caravale and Tosato (1980) and Caravale (1985) further emphasized the problematic nature of the

convergence process. Much of the complexity of this process arises from the intertwined dynamics of distributional change and population growth typical of the Ricardian system. Day (1983) has shown that characterization of the population dynamics by itself may be sufficient to generate extremely erratic or 'chaotic' motions. Bhaduri and Harris (1987) analyse the essential dynamics of the Ricardian system as it is governed solely by the interplay of distribution and accumulation in a model similar to the present one. They find that the model can generate very complex 'chaotic' movements instead of any smooth and gradual convergence to the stationary state. The possibility of such behaviour is shown to depend uniquely on the initial configuration of parameters. This result should lead one to question the presumption that the Ricardian system necessarily converges to a stationary state.

The Malthusian Population Dynamics

A crucial role is played in the classical analysis by the population dynamics deriving from the Malthusian law of population growth. In particular this law requires that population grows in response to a rise of wages above subsistence. This response mechanism is supposed to provide the labour requirements for expansion and thereby hold wages in check. But this is evidently a highly implausible principle on which to base an account of the process of capitalist expansion. If capitalism had to depend for its labour supply entirely upon such a demographic–biological response, it seems doubtful that sustained high rates of accumulation could continue for long or even that accumulation could ever get started. This is because, first, there must exist a biological upper limit to population expansion. Accumulation at rates above this limit would drive up the wage to such a level as to reduce or perhaps choke off the possibility of continued accumulation. For the classical labour supply principle to work, it must be presumed arbitrarily that this limit is sufficiently far out or, equivalently, that the supply curve is sufficiently elastic over a wide range.

Even if it is granted that population growth is significantly responsive to the level of wages, it is still the case that the adjustment of population is inherently a long drawn-out process having only a negligible effect on the actual labour supply in any short period of time. In the interim, any sizeable spurt of accumulation must then cause wages to be bid up, eat into profits, and bring accumulation itself, to a halt. From the start, therefore, accumulation could never get going in such a system. Even if it did, its continuation would always be in jeopardy because the mechanism of adjustment of labour supply is an inherently unreliable one, fraught with the possibility that at any time wages may rise to eat up the profits that are the wellspring of accumulation.

This feature of classical analysis was soundly criticized and rejected by Marx (1867, pp. 637–9). In its place, he sought to introduce a principle that was internal to the accumulation process, which would account for the continuing generation of a supply of labour to meet the needs of accumulation from within the accumulation process itself. This was the principle of the reserve army of labour or the 'law of relative surplus population' (1897, ch. 25, Sections 3 and 4). The reserve army results from a process of 'recycling' of labour through its displacement from existing employment due to mechanization and structural changes in production. In addition to this pool of labour there are other possible sources of increased labour supply to feed the accumulation process. These originate, for instance, in increased labour force participation rates among existing workers, in labour migration, and in the erosion of household work and other forms of non-capitalist production. Capital export to other regions can play the same role. These sources have been observed historically to be more or less significant at various times and places. It appears, therefore, that there is considerable flexibility of labour supply, and hence of accumulation, even without taking account of population growth. The existence of population growth certainly adds to the pool of available labour, as is now widely recognized. But the singular and unique role attributed to it by the Malthusian theory has by now been discredited and abandoned.

Conclusion

The classical economists are often regarded as 'pessimistic' in their prognosis for economic growth. It is said that they constituted economics as the 'dismal science'. Still, there is much to be learned, that is of contemporary relevance, from a close examination of their analytical system. What emerges from such an examination is a complex structure of ideas expressing a deep understanding of the nature of capitalism as an economic system, the sources of its expansionary drive, and the barriers or limits to its expansion. Their ideas were essentially limited, however, to the conditions of a predominantly agrarian economy, without significant change in methods of production, in which, because of the limited quantity and diminishing fertility of the soil, growth is arrested by increasing costs of production of agricultural commodities. Their analysis underestimated the far-reaching character of technological change as a powerful and continuing force in transforming the conditions of productivity both in agriculture and in industry. While they clearly perceived the possibilities opened up by international trade and foreign investment, they failed to incorporate these elements as integral components of a systematic theory of the growth process. It remained for Marx to pinpoint some of the major limitations and deficiencies of the classical analysis and to develop an analysis of the capitalist accumulation process that went beyond that of the classical economists in many respects while also leaving many unresolved questions. Subsequent work has continued to address the issues with limited success. Still today, the theory of growth of capitalist economies continues to be one of the most fascinating and still unresolved areas of economic theory.

See Also

- ▶ Development Economics
- ▶ Profit and Profit Theory
- ▶ Ricardo, David (1772–1823)
- ▶ Surplus

Bibliography

Bhaduri, A., and D.J. Harris. 1987. The complex dynamics of the simple Ricardian system. *Quarterly Journal of Economics* 102: 893–901.

Caravale, G.A., ed. 1985. *The legacy of Ricardo*. Oxford: Blackwell.

Caravale, G.A., and D.A. Tosato. 1980. *Ricardo and the theory of value, distribution and growth*. London: Routledge & Kegan Paul.

Casarosa, C. 1978. A new formulation of the Ricardian system. *Oxford Economic Papers* 30: 38–63.

Day, R.H. 1983. The emergence of chaos from classical economic growth. *Quarterly Journal of Economics* 98: 201–213.

Gordon, K. 1983. Hicks and Hollander on Ricardo: A mathematical note. *Quarterly Journal of Economics* 98: 721–726.

Harris, D.J. 1978. *Capital accumulation and income distribution*. Stanford: Stanford University Press.

Harris, D.J. 1981. Profits, productivity, and thrift: The neoclassical theory of capital and distribution revisited. *Journal of Post Keynesian Economics* 3: 359–382.

Harris, D.J. 1988. On the classical theory of competition. *Cambridge Journal of Economics* 12: 139–167.

Hicks, J.R., and S. Hollander. 1977. Mr. Ricardo and the moderns. *Quarterly Journal of Economics* 91: 351–369.

Kaldor, N. 1956. Alternative theories of distribution. *Review of Economic Studies* 23: 83–100.

Keynes, J.M. 1936. *The general theory of employment, interest, and money*. New York: Harcourt, Brace.

Lewis, W.A. 1954. Economic development with unlimited supplies of labour. *The Manchester School* 22 (2): 139–191.

Malthus, T.R. 1798. *Essay on the principle of population*. 1st ed, 1926. London: Macmillan.

Malthus, T.R. 1820. *Principles of political economy*. Reprinted in *The works and correspondence of David Ricardo*, vol. II, ed. P. Sraffa and M. Dobb. Cambridge: Cambridge University Press, 1951.

Marx, K. 1867. *Capital*. Vol. 1. New York: International Publishers.

Marx, K. 1894. *Capital*. Vol. 3. New York: International Publishers.

Marx, K. 1973. *Grundrisse*. Harmondsworth: Penguin Books.

Meek, R.L. 1967. *Economics and ideology and other essays*. London: Chapman & Hall.

Pasinetti, L. 1960. A mathematical formulation of the Ricardian system. *Review of Economic Studies* 27 (2): 78–98.

Pasinetti, L. 1977. *Lectures on theory of production*. New York: Columbia University Press.

Ricardo, D. 1951–1973. *The works and correspondence of David Ricardo,* vols. I–XI, ed. P. Sraffa with the collaboration of M.H. Dobb. Cambridge: Cambridge University Press.

Samuelson, P. 1978. The canonical classical model of political economy. *Journal of Economic Literature* 16: 1415–1434.
Schumpeter, J. 1934. *The theory of economic development*. New York: Oxford University Press.
Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations*. New York: Modern Library.
Sraffa, P. 1960. *Production of commodities by means of commodities*. Cambridge: Cambridge University Press.
Tucker, G. 1960. *Progress and profits in British economic thought 1650–1850*. Cambridge: Cambridge University Press.

Classical Growth Models

Donald J. Harris

Analysis of the process of economic growth was a central feature of the work of the English classical economists, as represented chiefly by Adam Smith, Thomas Malthus and David Ricardo. Despite the speculations of others before them, they must be regarded as the main precursors of modern growth theory. The ideas of this school reached their highest level of development in the works of Ricardo.

The interest of these economists in problems of economic growth was rooted in the concrete conditions of their time. Specifically, they were confronted with the facts of economic and social changes taking place in contemporary English society as well as in previous historical periods. Living in the 18th and 19th centuries, on the eve or in the full throes of the industrial revolution, they could hardly help but be impressed by such changes. They undertook their investigations against the background of the emergence of what was to be regarded as a new economic system – the system of industrial capitalism. Political economy represented a conscious effort on their part to develop a scientific explanation of the forces governing the operation of the economic system, of the actual processes involved in the observed changes that were going on, and of the long-run tendencies and outcomes to which they were leading.

The interest of the classical economists in economic growth derived also from a philosophical concern with the possibilities of 'progress', an essential condition of which was seen to be the development of the material basis of society. Accordingly, it was felt that the purpose of analysis was to identify the forces in society that promoted or hindered this development, and hence progress, and consequently to provide a basis for policy and action to influence those forces. Ricardo's campaign against the Corn Laws must obviously be seen in this light, as also Malthus's concern with the problem of population growth and Smith's attacks against the monopoly privileges associated with mercantilism.

Of course, for these economists, Smith especially, progress was seen from the point of view of the growth of national wealth. Hence, the principle of national advantage was regarded as an essential criterion of economic policy. Progress was conceived also within the framework of a need to preserve private property and hence the interests of the property-owning class. From this perspective, they endeavoured to show that the exercise of individual initiative under freely competitive conditions to promote individual ends would produce results beneficial to society as a whole. Conflicting economic interests of different groups could be reconciled by the operation of competitive market forces and by the limited activity of 'responsible' government.

As a result of their work in economic analysis the classical economists were able to provide an account of the broad forces that influence economic growth and of the mechanisms underlying the growth process. An important achievement was their recognition that the accumulation and productive investment of a part of the social product is the main driving force behind economic growth and that, under capitalism, this takes the form mainly of the reinvestment of profits. Armed with this recognition, their critique of feudal society was based on the observation among others,

Owing to an error on the part of Springer the content of this chapter has been published twice. This duplicate has been retracted.

that a large part of the social product was not so invested but was consumed unproductively.

The explanation of the forces underlying the accumulation process was seen as the heart of the problem of economic growth. Associated with accumulation is technical change as expressed in the division of labour and changes in methods of production. Smith, in particular, placed heavy emphasis on the process of extension of division of labour, but there is, in general, no systematic treatment of the relation between capital accumulation and technical change in the work of the classical economists. It later becomes a pivotal theme in the work of Marx and is subjected there to detailed analysis (see, for instance, *Capital,* I, part 4). To these basic forces in economic growth they added the increase in the supply of labour available for production through growth of population. Their analysis of the operation of these forces led them to the common view, though they quite clearly differed about the particular causes, that the process of economic growth under the conditions they identified raises obstacles in its own path and is ultimately retarded, ending in a state of stagnation – the 'stationary state'.

The conception of the stationary state as the ultimate end of the process of economic growth is often interpreted as a 'prediction' of the actual course of economic development in 19th-century England. There is no doubt that it was for a time so regarded by some, if not all, of the economists and their contemporaries, though the weight that was assigned to this particular aspect of the conception by Ricardo himself is a matter of some dispute. What is more significant, however, is that this conception served to point to a particular social group, the landlord class, who benefited from the social product without contributing either to its formation or to 'progress' and who, by their support of the corn laws and associated restrictions on foreign trade, acted as an obstacle to the only effective escape from the path to a stationary state, that is, through foreign trade.

In examining the work of the classical economists we find also that problems of economic growth were analysed through the application of general economic principles, viewing the economic system as a whole, rather than in terms of a separate theory of economic growth as such. These principles were such as to recognize basic patterns of interdependence in the economic system and interrelatedness of the phenomena of production, exchange, distribution, and accumulation. In sum, what we find in classical economic analysis is a necessary interconnection between the analysis of value, distribution, and growth. Because of these interconnections it was by no means possible to draw a sharp dividing line between the inquiry into economic growth and that into other areas of political economy. As Meek (1967, p. 187) notes:

> To Smith and Ricardo, the macroeconomic problem of the 'laws of motion' of capitalism appeared as the primary problem on the agenda, and it seemed necessary that the whole of economic analysis – including the basic theories of value and distribution – should be deliberately oriented towards its solution.

Distribution of the social product was seen to be connected in a definite way with the performance of labour in production and with the pattern of ownership of the means of production. In this regard, Labour, Land, and Capital were distinguished as social categories corresponding to the prevailing class relationships among individuals in contemporary society: the class of labourers consisted of those who performed labour services, landlords were those who owned titles or property in land, and capitalists were those who owned property in capital consisting of the sum of exchangeable value tied up in means of production and in the 'advances' which go to maintain the labourers during the production period. Each class received income or a share in the product according to specified rules: for the owners, the rule was based on the total amount of property which they owned – so much rent per unit of land, so much profit per unit of capital (and, for the class of finance capitalists or 'rentiers' who lent money at interest, so much interest per unit of money lent). For labourers it was based on the quantity of labour services performed: so much wages per hour.

Accumulation and distribution were seen to be interconnected through the use that was made by

different social classes of their share in the product. Basic to this view was a conception, taken over from the Physiocrats, of the social surplus as that part of the social product which remained after deducting the 'necessary costs' of production consisting of the means of production used up and the wage goods required to sustain the labourers employed in producing the social product. This surplus was distributed as profits, interest, and rent to the corresponding classes of property owners. For the classical economists, the possibility of accumulation was governed by the size and mode of utilization of this surplus. Accordingly, their analysis placed emphasis upon those aspects of distribution and of the associated class behaviour which had a direct connection with the disposal of the surplus and therefore with growth. In particular, it was assumed that, typically, workers consumed their wages for subsistence, capitalists reinvested their profits and landlords spent their rents on 'riotous living'. On the other side, accumulation would also influence the distribution of income as the economy expanded over time.

It was this absolutely strategic role of the size and use of the surplus, viewed from the perspective of the economy as a whole and of its process of expansion, which dictated the significance of the distribution of income for classical economic analysis. Thus, for Ricardo especially, investigation of the laws governing distribution became the focus of analysis. In a letter to Malthus, Ricardo wrote (*Works,* VIII, pp. 278–9): 'Political Economy you think is an inquiry into the nature and causes of wealth; I think it should rather be called an inquiry into the laws which determine the division of the produce of industry among the classes which occur in its formation.' What was of crucial significance in this connection was the rate of profits because of its connection with accumulation, both as the source of investment funds and as the stimulus to further investment.

Having 'got rid of rent' as the difference between the product on marginal land and that on intra-marginal units, the Ricardian analysis focused on profits as the residual component of the surplus. Under the simplifying conditions on which the analysis was constructed, there emerged a very clear and simple relationship between the wage rate and the overall rate of profits, determined within a single sector of the economy – the corn-producing sector. The special feature of corn as a commodity was that it could serve both as capital good (seed corn) in its own production and as wage good to be advanced to the workers. With the wage rate fixed in terms of corn, the rate of profit in corn production is uniquely determined as the ratio of net output of corn per man minus the wage to the sum of capital per man consisting of seed corn and the fund of corn as wage good. Competition ensures that the same rate of profit enters into the price of all other commodities that are produced with indirect labour. The overall rate of profits, determined in this way, varies inversely with the corn wage. But, as soon as it is recognized that the wage and/or the capital goods employed in corn production consist of other commodities besides corn, the rate of profits can no longer be determined in this way. For the magnitude of the wage and of the total capital then depends on the prices of those commodities, and these prices incorporate the rate of profit. Attention then has to be directed to explaining the rate of profit by taking account of the whole system of prices. For this purpose the theory of value is called upon to provide a solution and Ricardo struggled with this problem until the end of his life. An elegant solution has now been worked out by Sraffa (1960) which shows that, in a system of many produced commodities, with the real wage rate given at a specified level, the rate of profit is determined by the given wage and the conditions of production of the commodities that are 'basics'. It so happens that Ricardo's case of corn is just such a 'basic' commodity in the strict sense that it enters directly and indirectly into the production of every commodity including itself.

Given the perceived centrality of the rate of profit in a capitalist economy, for classical political economy it becomes a crucial problem in the theory of economic growth to account for movements in the rate of profit associated with the process of capital accumulation and development of the economy. Such movements are a decisive reference point for understanding the long-term evolution of the economy. The classical answer to this problem, as worked out most coherently by

Ricardo, is that in a closed economy there is an inevitable tendency for the rate of profit to fall in the course of the accumulation process and, hence, that the accumulation process itself is brought to a halt by its own logic.

Marx was later to propose this falling tendency of the rate of profit (FTRP) as a *law.* He considered it to be 'the most important law of modern political economy' (*Grundrisse,* p. 748; *Capital,* III, part 3). He was, of course, following in the tradition of the classical economists in which the same idea had been firmly entrenched, though supported on different grounds. But, interestingly enough, it is also the case that there exists a distinct conception of a FTRP within neoclassical theory (see Harris 1978, ch. 9; 1981). In Keynes, as well, the idea is embodied in his projection of the long-term prospects for capitalism resulting in the 'euthanasia of the rentier' (1936, pp. 375–6). In Schumpeter (1934), it occurs in the form of the idea that the profitability of innovations tends inevitably to be eroded so that the economy settles back to the conditions of the 'circular flow' in the absence of new innovations. Though it is based in each case on quite different foundations, this conception is one of the most striking and persistent uniformities across different schools of economic thought. (For a discussion of the long history of the idea of a falling rate of profit, see Tucker 1960).

A Model of Accumulation

The essential features of the classical argument regarding the accumulation process can be exhibited with a simple model adapted from Kaldor (1956) and Pasinetti (1960). This model formalizes the Ricardian conception of an agricultural economy producing a single product, 'corn', under capitalist conditions. Land is of differing fertility and labour is applied in fixed proportion to less and less fertile land. Accordingly, the average and marginal product of labour falls as the margin of cultivation is extended through capital accumulation and increase of employment on the land. The system may indifferently be assumed to expand on the extensive or intensive margins of available land. Also, it does not matter for this analysis that there exists any production outside agriculture. It would turn out, in any case, that the overall average rate of profit for the economy as a whole is determined by the agricultural rate of profit or, in the general case, by the conditions of production of 'basics' (cf. Sraffa 1960; Pasinetti 1977). Of course, in a system with many produced commodities, it is not possible to define 'less fertile land' independently of the rate of profit (Sraffa 1960). However, this problem does not arise in this simplified model of a corn-producing economy. We deliberately abstract from complications associated with the Malthusian population dynamics. This is perhaps the most problematic feature of the classical conception and we return to it below. Meanwhile, it is simply assumed, as in Lewis (1954), that a labour force is in perfectly elastic supply at some conventionally fixed real wage rate equal to 'subsistence'.

Let the production function relating output Y to labour input L be

$$\begin{aligned} Y = F(L) \qquad & F(0) \geqq 0 \\ & F' > w^* > 0 \\ & F'' < 0 \end{aligned} \tag{1}$$

which satisfies the law of diminishing returns and allows for the existence of a surplus product above the 'subsistence' wage-rate w^*. Total capital K consists entirely of wages W (the 'wage fund') advanced at the beginning of the production period to hire labour. Thus

$$K = W = wL \tag{2}$$

We are here, for simplicity, neglecting capital as seed-corn and inputs of fixed capital are ignored. Total output is distributed between payment of rent R to landlords, profits P to capitalists, and replacement of the wage fund:

$$Y = R + P + W \tag{3}$$

Given the margin of cultivation reached at any time, the level of land rent is determined as the difference between the average and marginal

product of labour at the prevailing level of employment:

$$R = \left(\frac{F(L)}{L} - F'\right)L \tag{4}$$

Profit emerges as the residual

$$P = (F' - w^*)L \tag{5}$$

It follows that the rate of profit r is determined from

$$r = \frac{P}{W} = \frac{F'}{w^*} - 1 \tag{6}$$

It is the dynamics of the wage fund which represents the process of accumulation in this model. Accumulation of capital consists of the growth of the wage fund with a corresponding increase of employment. Additions to the wage fund come entirely from investment of capitalists' profits since the spend-thrift landlords consume their share of the surplus. If the capitalists invest a proportion of profits equal to a, then

$$\Delta W = \alpha P \qquad 0 < \alpha < 1 \tag{7}$$

The proportion α need not be a constant. It could vary in a manner dependent on the rate of profit as suggested by Ricardo's idea that

[the capitalists'] motive for accumulation will diminish with every diminution of profit, and will cease altogether when their profits are so low as not to afford them an adequate compensation for their trouble and the risk which they must necessarily encounter in employing their capital productively (*Works,* I, p. 122).

In that case we have

$$\begin{aligned} \alpha &= \alpha(r) \qquad \alpha' > 0 \\ \alpha(r^*) &= 0 \end{aligned} \tag{8}$$

where r^* is the capitalists' minimum acceptable rate of profit. By definition the rate of capital accumulation is $g = \Delta W/W$, and from (6), (7), and (8) it follows that

$$g = \alpha(r) \cdot r \tag{9}$$

Thus, the rate of accumulation is uniquely dependent on the profit rate.

The movement in the profit rate as accumulation proceeds can be derived from (6). Evidently, as employment increases the marginal product of labour falls. The rate of profit must therefore fall. It continues to fall as long as there is any increment to the wage fund so as to employ extra labour on the available land. The process comes to a half when the profit rate is so low that accumulation ceases. The economy is then at the stationary state.

In this model, the capitalists are caught between, on the one hand, the diminishing productivity of labour as the margin of cultivation is extended and, on the other, the need to pay the ongoing wage rate in order to secure labour for employment. As the productivity of labour falls on the marginal land the pressure of land rent increases for the existing intra-marginal units. The capitalists must therefore pay out an increasing share of the surplus to the landlords. In this way they gradually lose command over the investible surplus of the economy to the landlord class. This distributional conflict between the landlord class and the capitalists constitutes a central feature of the process that drives the economy towards its ultimate stationarity. The impenetrable barrier in the process is the diminishing fertility of the soil. More generally, it is the limitation of natural resources, in this case land, which brings the process to a halt. In this respect the classical model is a particular case of resource-limited growth. Any other limited resource would have the same effect, through increasing 'rents' for that resource. At the same time, this consequence is also the product of the capitalists' own actions in relentlessly seeking to expand the size of their capital.

The underlying dynamic process which expresses this conflictive evolution of capitalist accumulation has usually been assumed in the literature to converge towards the stationary state (cf. Pasinetti 1960; Samuelson 1978). Some reservation on this question of convergence was originally expressed by Hicks and Hollander

(1977) and followed up by Gordon (1983). Subsequent discussion by Casarosa (1978), Caravale and Tosato (1980) and Caravale (1985) further emphasized the problematic nature of the convergence process. Much of the complexity of this process arises from the intertwined dynamics of distributional change and population growth typical of the Ricardian system. Day (1983) has shown that characterization of the population dynamics by itself may be sufficient to generate extremely erratic or 'chaotic' motions. In a recent paper, Bhaduri and Harris (1986) analyse the essential dynamics of the Ricardian system as it is governed solely by the interplay of distribution and accumulation in a model similar to the present one. They find that the model can generate very complex 'chaotic' movements instead of any smooth and gradual convergence to the stationary state. The possibility of such behaviour is shown to depend uniquely on the initial configuration of parameters. This result should lead one to question the presumption that the Ricardian system necessarily converges to a stationary state.

The Malthusian Population Dynamics

A crucial role is played in the classical analysis by the population dynamics deriving from the Malthusian Law of Population Growth. In particular this law requires that population grows in response to a rise of wages above subsistence. This response mechanism is supposed to provide the labour requirements for expansion and thereby hold wages in check. But this is evidently a highly implausible principle on which to base an account of the process of capitalist expansion. If capitalism had to depend for its labour supply entirely upon such a demographic-biological response, it seems doubtful that sustained high rates of accumulation could continue for long or even that accumulation could ever get started. This is because, first, there must exist a biological upper limit to population expansion. Accumulation at rates above this limit would drive up the wage to such a level as to reduce or perhaps choke off the possibility of continued accumulation. For the classical labour supply principle to work it must be presumed arbitrarily that this limit is sufficiently far out or, equivalently, that the supply curve is sufficiently elastic over a wide range.

Even if it is granted that population growth is significantly responsive to the level of wages, it is still the case that the adjustment of population is inherently a long drawn-out process having only a negligible effect on the actual labour supply in any short period of time. In the interim, any sizeable spurt of accumulation must then cause wages to be bid up, eat into profits, and bring accumulation itself, to a halt. From the start, therefore, accumulation could never get going in such a system. Even if it did, its continuation would always be in jeopardy because the mechanism of adjustment of labour supply is an inherently unreliable one, fraught with the possibility that at any time wages may rise to eat up the profits that are the wellspring of accumulation.

This feature of classical analysis was soundly criticized and rejected by Marx (*Capital*, I, pp. 637–9). In its place, he sought to introduce a principle that was internal to the accumulation process, that would account for the continuing generation of a supply of labour to meet the needs of accumulation from within the accumulation process itself. This was the principle of the reserve army of labour or the 'law of relative surplus population' (*Capital,* I, ch. 25, sections 3 and 4). The reserve army results from a process of 'recycling' of labour through its displacement from existing employment due to mechanization and structural changes in production. In addition to this pool of labour there are other possible sources of increased labour supply to feed the accumulation process. These originate, for instance, in increased labour force participation rates among existing workers, in labour migration, and in the erosion of household work and other forms of non-capitalist production. Capital export to other regions can play the same role. These sources have been observed historically to be more or less significant at various times and places. It appears, therefore, that there is considerable flexibility of labour supply, and hence of accumulation, even without taking account of

population growth. The existence of population growth certainly adds to the pool of available labour, as is now widely recognized. But the singular and unique role attributed to it by the Malthusian theory has by now been discredited and abandoned.

Conclusion

The Classical economists are often regarded as 'pessimistic' in their prognosis for economic growth. It is said that they constituted economics as the 'dismal science'. Still, there is much to be learned, that is of contemporary relevance, from a close examination of their analytical system. What emerges from such an examination is a complex structure of ideas expressing a deep understanding of the nature of capitalism as an economic system, the sources of its expansionary drive, and the barriers or limits to its expansion. Their ideas were essentially limited, however, to the conditions of a predominantly agrarian economy, without significant change in methods of production, in which, because of the limited quantity and diminishing fertility of the soil, growth is arrested by increasing costs of production of agricultural commodities. Their analysis underestimated the far-reaching character of technological change as a powerful and continuing force in transforming the conditions of productivity both in agriculture and in industry. While they clearly perceived the possibilities opened up by international trade and foreign investment, they failed to incorporate these elements as integral components of a systematic theory of the growth process. It remained for Marx to pinpoint some of the major limitations and deficiencies of the classical analysis and to develop an analysis of the capitalist accumulation process that went beyond that of the classical economists in many respects while also leaving many unresolved questions. Subsequent work has continued to address the issues with limited success. Until today, the theory of growth of capitalist economies continues to be one of the most fascinating and still unresolved areas of economic theory.

See Also

- ▶ British Classical Economics
- ▶ Classical Distribution Theories
- ▶ Smith, Adam (1723–1790)

C

Bibliography

Bhaduri, A., and D.J. Harris. 1986. The complex dynamics of the simple Ricardian system. *Quarterly Journal of Economics* 102: 893.

Caravale, G.A. (ed.). 1985. *The legacy of Ricardo*. Oxford: Blackwell.

Caravale, G.A., and D.A. Tosato. 1980. *Ricardo and the theory of value, distribution and growth*. London: Routledge & Kegan Paul.

Casarosa, C. 1978. A new formulation of the Ricardian system. *Oxford Economic Papers* 30(1): 38–63.

Day, R.H. 1983. The emergence of chaos from classical economic growth. *Quarterly Journal of Economics* 98(2): 201–213.

Gordon, K. 1983. Hicks and Hollander on Ricardo: A mathematical note. *Quarterly Journal of Economics* 98(4): 721–726.

Harris, D.J. 1978. *Capital accumulation and income distribution*. Stanford: Stanford University Press.

Harris, D.J. 1981. Profits, productivity, and thrift: The neoclassical theory of capital and distribution revisited. *Journal of Post Keynesian Economics* 3(3): 359–382.

Hicks, J.R., and S. Hollander. 1977. Mr. Ricardo and the moderns. *Quarterly Journal of Economics* 91(3): 351–369.

Kaldor, N. 1956. Alternative theories of distribution. *Review of Economic Studies* 23: 83–100.

Keynes, J.M. 1936. *The general theory of employment, interest, and money*. New York: Harcourt, Brace.

Lewis, W.A. 1954. Economic development with unlimited supplies of labour. *The Manchester School* 22: 139–191.

Malthus, T.R. 1798. *Essay on the principle of population*, 1st ed. London: Macmillan, 1926.

Malthus, T.R. 1820. Principles of political economy. Reprinted in The works and correspondence of David Ricardo, ed. P. Sraffa and M. Dobb, Vol. II. Cambridge: Cambridge University Press, 1951.

Marx, K. 1867. *Capital*, vol. I. New York: International Publishers, 1967.

Marx, K. 1973. *Grundrisse*. Harmondsworth: Penguin Books.

Meek, R.L. 1967. *Economics and ideology and other essays*. London: Chapman & Hall.

Pasinetti, L. 1960. A mathematical formulation of the Ricardian system. *Review of Economic Studies* 27(2): 78–98.

Pasinetti, L. 1977. *Lectures on the theory of production*. New York: Columbia University Press.

Ricardo, D. 1951–73. *The works and correspondence of David Ricardo.* Ed. P. Sraffa with the collaboration of M. H. Dobb. Cambridge: Cambridge University Press.
Samuelson, P. 1978. The canonical classical model of political economy. *Journal of Economic Literature* 16: 1415–1434.
Schumpeter, J. 1934. *The theory of economic development*. New York: Oxford University Press.
Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations*. New York: Modern Library, 1937.
Sraffa, P. 1960. *Production of commodities by means of commodities*. Cambridge: Cambridge University Press.
Tucker, G. 1960. *Progress and profits in British economic thought 1650–1850*. Cambridge: Cambridge University Press.

Classical Production Theories

Giorgio Gilibert

Keywords

Advances; Agriculture; Babbage, C.; Capital accumulation; Circular flow; Classical economics; Classical production theories; Division of labour; Factory system; Increasing returns; Industrial Revolution; Marginal revolution; Marx, K. H.; Net product; Physiocracy; Production; Productive and Unproductive Inputs; Quesnay, F.; Rent; Ricardo, D.; Smith, A.; Stationary State; Surplus; Technical change; Tugan-Baranovsky, M. I.; Ure, W.

JEL Classifications

D2

A theory of production cannot be said to have existed before the middle of the 18th century. The very word production was previously used in its narrow etymological sense (from the Latin *producere*, to bring forth) of giving birth to new material objects; and it was therefore normally confined to the fruits of the earth. 'When we speak of it', writes Daniel Defoe, 'as the Effect of Nature 'tis *Product* or Produce; when as the Effect of Labour 'tis *Manufacture*' (1728, p. 1).

It is with the writings of the French *économistes* that the term receives a precise technical meaning. At first sight, the Physiocratic terminology is not particularly novel: the words production, productivity, and so on are carefully reserved for agriculture; manufacture, as a mere transforming activity, is considered as eminently sterile. But Quesnay's fundamental innovation lies in the theory behind the terminology: it is not (or not so much) because of some physical property that agriculture is said to be productive, but because it is the only activity capable of generating a net revenue (rent). The way was, however, paved for the recognition of the productivity of non-agricultural activities, provided that the peculiar assumption of rent as the only possible net revenue was dropped (that is, that profit was accepted as a legitimate form of net revenue). This step was taken, a few years later, by Adam Smith.

In the following decades, production became one of the main topics of political economy; this was later sanctioned by the standard structure adopted by economic textbooks, whose first section typically came to be devoted to production. The first English example of this arrangement is given by the *Elements of Political Economy* published by James Mill in 1821 (following in this respect in Say's steps), the same year in which Robert Torrens brought out his *Essay on the Production of Wealth*. Eventually, in Marxian economics, production analysis achieved the status of a cornerstone of the whole theory of social change.

In the second half of the 19th century, as a consequence of the so-called marginalist revolution, the focus of economic theory tended to shift from the sphere of production to that of exchange. Production theory was squeezed into the general framework of the optimal allocation of scarce resources: a framework originally developed to deal with the problem of pure exchange. The theory originally devised by Quesnay seemed, about one century after its birth, to conclude its own theoretical lifetime.

François Quesnay was the first to analyse the system of production and consumption as a single complex process. He looked for the 'natural laws' by which it was regulated, laws which were

independent of the will of man but discoverable by the light of reason. The attempt to present the interplay of these laws in an abstract and manageable way originated the first theoretical model of the history of economic analysis.

The Physiocratic doctrine presents, though often under a misleading feudal disguise, most of the leading ideas of the classical theory of capitalist production. First and foremost, the picture of the system of production and consumption as a circular process: no one will ever deny that consumption is the ultimate end of production, but it is essential to bear in mind the simple fact that past production determines present consumption, and that consumption in turn is nothing but the necessary condition for future production.

The idea of production as a circular process immediately suggests the notion of surplus: if the economy produces more than the minimum necessary for the process to be repeated, then there is a surplus. Its value was called 'net product' by Quesnay: this is the strategic variable for economic activity. The nations' prosperity can be assessed according to the size of their annual net product.

The answers given by the Physiocrats to the fundamental questions of the origins and destination of the net product account for their peculiar class analysis. They assumed that a net product was yielded exclusively by land-using activities; that is, that revenues could be higher than costs only in agriculture, and therefore rent was the only conceivable net revenue. The class of those engaged in agriculture (farmers, the labourers being equated to cattle) was thus called 'productive', in contrast to the 'sterile' class of those engaged in manufacture (artisans); the class of landowners got the whole net product, under the form of rent.

Since the process of production takes time (the agricultural year) it requires advances: for instance, the labourers' subsistence must be available before the harvest. Quesnay distinguishes between annual advances (working capital: seed, subsistence), which are wholly used up in the course of the production process, and original advances (fixed capital, for which a depreciation is allowed), which are not. It is perhaps worth noticing that the word capital was commonly in use in the economic literature of the 18th century. Quesnay's unusual terminology was presumably due to his intention of stressing the physical nature of the advances required by the production process, as opposed to the current meaning of capital as a sum of money employed in trade.

The characteristic agricultural bias of the Physiocrats is shown not only by their doctrine of the sterility of manufacture, but also by the essentially static nature of their models. If the economy is organized according to the natural order, that is according to the 'evident' laws discovered by the economists, it will rapidly attain the maximum level of output consistent with the country's amount of arable land and with the state of technology. Indeed, the *Tableaux* depict this prosperous and stationary situation.

Both these aspects are definitely abandoned by Adam Smith. Precisely because production takes time, and wages, materials and equipment have to be anticipated, the owners of these advances, the capitalists, are naturally entitled to a part of the net revenue, the profits. The advances are consumed by productive workers or used up as raw materials and wear and tear of equipment; the return, in manufacture as well as in agriculture, will normally cover their cost with an addition, which constitutes the profit.

The Smithian capitalist is thrifty and industrious; his profits are well above subsistence, and he will normally save most of them and employ these savings as capital, in order to obtain an additional profit in the future. As a result of these decisions, the capital of the nation as a whole, the fund that sets productive labour to work for the purpose of profit, naturally tends to increase each year in the course of economic progress.

In this way, Smith gave a clear-cut answer to an old dilemma. In his century, two traditional ideas coexisted unreconciled side by side: on the one hand, by analogy with the behaviour of a good husband, thriftiness was praised as a social virtue; on the other hand, it was maintained that a buoyant consumption stimulated trade. In Smith's view, every frugal man is a benefactor, every prodigal man a 'public enemy'.

The progressive state of the economy – it is written in the *Wealth of Nations* – 'is in reality the chearful and the hearty state to all the different orders of the society. The stationary is dull; the declining, melancholy' (Smith 1776, p. 99). The analysis is here primarily concerned with the process of capital accumulation and is therefore necessarily dynamic.

The analysis of the accumulation of wealth inevitably involved the question of the final outcome of the process. It was a common belief – among classical economists – that the economy would eventually tend towards a stationary state. This could be seen optimistically as 'a full complement of riches' (Smith) or, on the contrary, as a sad motionless state (Ricardo); still, it could be considered as relatively far ahead in the future (Smith and, with a suitable economic policy, Ricardo) or just round the corner (J.S. Mill).

An interesting technical feature of the theory of production can be introduced in connection with this question. The advances of every industry are normally composed of commodities that are not produced by that industry. In other words, each industry must exchange part of its output on the market with the necessary inputs to start the production process again. These transactions, dictated by the technology in use, were clearly described by the *Tableau:* in this highly aggregate picture, the two activities considered, productive and sterile, are both essential to reproduction. But, in a more detailed framework, we can distinguish between those commodities which play a productive role as inputs, and those which do not ('luxuries'). The growth potential of the economy is affected only by the conditions of production of the first type of commodities ('basics' according to modern terminology).

Since every line of production requires labour, and workers consume food, foodstuffs are basics par excellence. Food production in turn requires land, a non-reproducible resource; the scarcity of land becomes therefore the limiting factor to accumulation. Land is essential, and is fixed in supply, so the eventual outcome of the growth process is the stationary state. (One might think that in this way we are back with the original Physiocratic perspective, but now attention is focused on the dynamic process rather than on its static outcome.)

David Ricardo presented a sophisticated version of this argument, in which the result that the growth process ends in a stationary state is analytically restated via his theory of profits. In evaluating this kind of argument, one must remember the vital *ceteris paribus* assumption, especially with regard to technology. Of course, the process of exhaustion of natural resources can be checked by improvements which affect agriculture. Ricardo has often been criticized for his allegedly cursory treatment of technical progress: one instance can be found in *The Logic of Political Economy* written a quarter of a century later (1844) by his follower Thomas de Quincey.

With Karl Marx, the concept of production acquires new and wider meanings; in a sense it leaves the narrow field of economic theory to become the cornerstone of a general theory of social systems and of history (the development of material production, notes Marx in the first book of *Capital*, is) 'the basis of any social life and of any true history'). The starting point of the analysis is the notion of production in its elementary form: men produce the necessaries for their existence; their productive activity is labour, which materializes into products. In other words, men produce the conditions for their material life. What men are is then determined by production; more specifically, by what is produced and by the way in which it is produced.

Production is essentially a social process: there are no 'natural laws' to be investigated, but social relations which are historically determined. These relations constitute the structure of society and determine its material and intellectual way of life. The evolution of religion, ethics, art and government is an ultimate consequences of the evolution of the social relations of production.

In his justly famous preface to the *Critique of Political Economy*, Marx has left a very effective summary statement of this approach:

> In the social production which men carry on they enter into definite relations that are independent of their will; these relations of production correspond to a definite stage of development of their material powers of production. The sum total of these

> relations of production constitutes the economic structure of society – the real foundation on which rise the legal and political superstructures and to which correspond definite forms of social consciousness. The mode of production in material life determines the general character of the social, political, and spiritual processes of life. It is not the consciousness of men that determines their existence, but, on the contrary, their social existence determines their consciousness. (Marx 1859, p. 100)

Production, distribution, exchange and consumption cannot be grasped in their essence but as successive moments of a unique circular process, thoroughly determined by the social conditions of production. Marx reproaches political economy for having arbitrarily separated the sphere of production, regulated by allegedly universal laws, from that of distribution, where we can take account of the social environment.

The search for universal laws of production has in turn led the economist to concentrate upon the trivial aspects of the phenomenon and to overlook the questions that are truly essential in investigating the present mode of production. For example, having defined as capital the set of the means of production, and having observed the obvious fact that men have always needed some kind of instrument to produce, the economists are ready to attribute a universal and ahistorical validity to the notion of capital. In this way, they have simply swept aside the key question: what is the socially determined relationship which turns an instrument used in production into 'capital'?

The formation of classical political economy historically coincided with the development of the factory system in manufacture. An early description of an integrated production process is offered by William Petty (1683) with reference to the watch trade. Another obvious reference is the famous pin factory described by Adam Smith in the first chapter of the *Wealth of Nations* (1776). In both cases, the division of labour is presented as the main virtue of the new form of productive organization: provided that the extent of the market is sufficient, it is maintained that output can be expanded more than proportionately with the labour employed in manufacture (increasing returns to scale).

Marx used these two examples to draw a distinction between the 'heterogeneous' manufacture (exemplified by Petty's watch-making activity) in which the final output is obtained by simple assemblage of 'partial and independent products', and the more sophisticated 'organic' manufacture (exemplified by Smith's pin factory) in which a series of successive operations gradually transforms the original raw material into the finished product.

Smith referred to three arguments in favour of the technical superiority of an ever increasing division of labour:

> first, to the increase of dexterity in every particular workman; secondly, to the saving of the time which is commonly lost in passing from one species of work to another; and lastly, to the invention of a great number of machines which facilitate and abridge labour, and enable one man to do the work of many. (Smith 1776, p. 17)

It has been observed that these arguments are not truly convincing. The importance attributed to increased dexterity conflicts with the relatively low level of skills required in contemporary factories (witness the common use of child labour). Time saving does not imply specialization by individuals: in principle, it could equally be attained by a suitable reorganization of the activity of a single artisan. And the introduction of machines does not seem to exhibit any necessary relation to the increasing division of tasks.

In fact the new organization of labour associated with the factory system did go along with the process of technical change associated with the Industrial Revolution. But its original role was primarily to discipline the manner in which the work was performed and to give the capitalist the power of controlling the production process in every single detail.

The introduction of machinery came after labour specialization and reinforced the need for a thorough organization of production. The effects of the introduction of the steam-engine and other complex machines were eventually studied by two scholars who possessed the necessary technical background, Charles Babbage (1832) and William Ure (1835); their tracts were very popular at the time and were widely used by the economists

(for example, by John Stuart Mill and Marx). They conceived of the control and management of a factory as that of a single complex machine, under the full control of the capitalist and with manual work brought to a minimum.

It is worth noticing that these speculations about the rational management of a highly mechanized factory were easily extended to society as a whole. At the turn of the century, Mikhail Tugan-Baranovsky (1905) dreamed of an economy in which machines were automatically produced by machines, and where the labour force was paradoxically reduced to one worker alone. In a similar vein, especially in Germany after the First World War, we find many suggestions for a 'rational' organization of the economy as if it were a giant *Konzern* (as an extreme example, see the 'natural economy' proposed by Otto Neurath (1921) for the ephemeral Bavarian republic).

Bibliography

Babbage, C. 1832. *On the economy of machine and manufactures*. London: Knight.

De Quincey, T. 1844. The logic of political economy. In *Collected writings*, ed. D. Masson, vol. 9. London: Black, 1897.

Defoe, D. 1728. *A plan of the English commerce*. Oxford: Blackwell, 1928.

Marx, K. 1859. Zur Kritik der politischen Ökonomie. In *Marx-Engels Gesamtausgabe*, vol. 2, pt. II. Berlin: Dietz, 1980.

Marx, K. 1867. Das Kapital, vol. 1. In *Marx-Engels Gesamtausgabe*, vol. 2, pt. V. Berlin: Dietz, 1983.

Mill, J. 1821. *Elements of Political Economy.* London: Baldwin.

Mill, J.S. 1848. *Principles of political economy.* Ed. J.-M. Robson. Toronto: University of Toronto Press, 1965.

Neurath, O. 1921. *Durch die Kriegswirtschaft zur Naturalwirtschaft*. Munich: Callway.

Petty, W. 1683. Another essay on political arithmetick. In *Economic writings of Sir William Petty*, ed. C.-H. Hull, vol. 2. Cambridge: Cambridge University Press, 1899.

Quesnay, F. 1759. *Tableau économique*. Ed. M. Kuczynski and R. Meek. London: Macmillan, 1972.

Ricardo, D. 1815. An essay on the influence of a low price of corn. In *The works and correspondence of David Ricardo*, ed. P. Sraffa, vol. 4. Cambridge: Cambridge University Press, 1951.

Ricardo, D. 1817. Principles of political economy. In *The works and correspondence of David Ricardo*, ed. P. Sraffa, vol. 1. Cambridge: Cambridge University Press, 1951b.

Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations*. Ed. R.H. Campbell, A.S. Skinner and W.B. Todd. Oxford: Clarendon Press, 1976.

Torrens, R. 1821. *An essay on the production of wealth*. London: Longman, Hurst, Rees, Orme & Brown.

Tugan-Baranovsky, M. 1905. *Theoretische Grundlagen des Marxismus*. German Trans. Leipzig: Duncker & Humblot.

Ure, A. 1835. *The philosophy of manufactures*. London: Knight.

Classroom Peer Effects

Jane Cooley

Abstract

A central objective of studies of peer effects in education production is to determine whether certain groupings of students can improve academic achievement. This article describes the challenges associated with identifying peer effects in education production and some solutions offered by the literature. We then review some of the existing evidence on peer spillovers and avenues for future research.

Keywords

Peer effects; Reflection problem; Tracking; Desegregation

JEL Classifications

I20; I21

Introduction

A central objective of studies of peer effects in education production is to determine whether certain groupings of students can improve academic achievement. For instance, do students benefit from being grouped with students of similar ability (academic tracking)? Do tracking policies

benefit higher ability students at the expense of lower ability students? Do mixed ability classrooms benefit lower ability students at the expense of those of higher ability? Do single-sex classrooms improve achievement for boys and girls? Does racial or socioeconomic integration improve achievement of traditionally disadvantaged students? Is racial integration more effective than socioeconomic integration at narrowing racial achievement gaps? What classroom groupings are most efficient?

Research that seeks to inform these policies can broadly be divided into two classes: event studies and achievement production studies. Event studies consider, for instance, actual events of desegregation (e.g. Guryan 2004; Harris 2007) or contrast academic tracking with mixed ability settings either internationally or across schools (see Gamoran 2009, for a recent overview). A key challenge for these studies is to separate a peer effect from disparities in difficult-to-measure resource inputs across settings. In the case of desegregation, black students may face lower resources or teacher quality prior to desegregation than in the integrated setting. In the case of tracking, teachers may teach differently in mixed ability than in tracked classes, or higher quality teachers may be assigned to higher ability tracks. Separating the effect of resources from peers can be important because it is possible, at least in principle, to reallocate resources without manipulating peer groups.

Another class of studies uses an achievement production framework and exploits variation in the composition of students across a set of classrooms, year groups, or schools to determine how different peer groupings affect student outcomes. In this case, instead of directly examining the policy, such as tracking or desegregation, these studies consider how higher percentages of nonwhite or higherachieving peers affect student performance. Like the event studies described above, achievement production studies also face the identification problem of separating peer effects from resources that may be correlated with observed student groupings. However, the key advantage of these studies is to begin to separate different sources of peer effects. For instance, if mixed ability classrooms are also more racially diverse, is it the ability of students or their racial composition that drives improvements in achievement? The attempt to disentangle channels of peer influence is fraught with its own set of challenges, as discussed below (see Cooley 2009).

As the literature on peer effects is extensive, evidence from only a small sample of recent studies is discussed. In particular, the focus is on achievement production types of studies. These are becoming increasingly common, particularly given new panel data sources on student achievement and peer groups, such as the state administrative data sets in the USA. We describe the identification challenges for these studies in more detail, some solutions offered in the literature, and the potential of these studies to inform policy.

Identification Challenges

Studies of peer effects in education production often take as a starting point a linearin-means production function. This provides a simple context to describe the basic challenges for identifying peer effects. Let Y_{igt} denote student i's achievement in peer group g at time period t. X_{it} denotes observed individual characteristics, which often include parental education, race, sex, and some measure of income. K_{gt} captures observed classroom inputs, such as teacher experience or expenditure, and μ_{gt} captures unobserved inputs, such as unobserved teacher quality. Achievement production is then

$$Y_{igt} = X_{it}\gamma_x + \overline{X}_{-igt}\gamma_{\overline{x}} + K_{gt}\gamma_k + \overline{Y}_{-igt}\gamma_{\overline{y}} + \mu_{gt} + \varepsilon_{igt}, \quad (1)$$

where peer spillovers derive from both mean peer characteristics $\overline{X}_{-igt}$ (contextual or exogenous effects) and mean peer achievement $\overline{Y}_{-igt}$ (the endogenous effect), using the language of Manski (1993). Often these specifications are estimated as valueadded models, conditioning on a student's lagged peer achievement to help control for prior inputs to achievement.

There are several challenges with identifying a causal effect of peers. First, unobservable shared effects (μ_{gt}) may be correlated with peer characteristics. This could occur through selection (nonrandom assignment) into classrooms or schools. It could also be the case that these unobservables vary systematically with the composition of the peer group if, for instance, teachers teach differently with different sets of students. Random assignment to peer groups would not eliminate the latter effect, and most peer effect studies implicitly attribute these types of reallocations in teacher effort to peers.

In this setting, random assignment is sufficient to recover a reduced form effect of peer characteristics. To see this, solve for average peer achievement and substitute back into equation (1). Then we have the reduced form equation for peer achievement:

$$Y_{igt} = \pi_0 + X_{it}\pi_x + \overline{X}_{-igt}\pi_{\bar{x}} + K_{gt}\pi_k + \pi_\mu \mu_{gt} + \zeta_{igt}, \quad (2)$$

where $\frac{\pi_{\bar{x}}=\gamma_{\bar{x}}(N-1)+(N-2)\gamma_{\bar{x}}\gamma_{\bar{y}}+\gamma_x\gamma_{\bar{y}}+\gamma_{\bar{x}}(N-2)}{(N-1)-(N-2)\gamma_{\bar{y}}-(N-1)\gamma_{\bar{y}}^2}$. Random assignment helps ensure that $\overline{X}$ is independent of μ_{gt}, which would permit the identification of the reduced form effect of peer characteristics, $\pi_{\bar{x}}$, given $E(\zeta_{igt}|X_{it}, \overline{X}_{-igt}, K_{gt}) = 0$. Using Manski (1993)'s terminology, $\pi_{\bar{x}} \neq 0$ means that *social effects* exist, in that either $\gamma_{\bar{x}} \neq 0$ and/or $\gamma_{\bar{y}} \neq 0$.

Second, even if students are randomly assigned to classrooms and peer characteristics satisfy some sort of independence with unobserved shared group inputs, there still exists the challenge of determining the causal mechanism of the peer effect. For instance, if ability is unobservable and correlated with observed characteristics X, it may not be possible to separate out an effect of the observed characteristic from unobservable ability.

Furthermore, it may not be possible to separate contextual ($\gamma_{\bar{x}}$) from endogenous ($\gamma_{\bar{y}}$) peer effects because of the simultaneity problem. Moffitt (2001) points out that what makes this simultaneity problem particularly difficult is the lack of compelling exclusion restrictions, a policy that shifts one student's achievement independently of his peers. Brock and Durlauf (2001a, b) make the important point that the identification challenge stems from the particularly restrictive functional form chosen in the linear-in-means model. In a more general nonlinear model, such as simply focusing on median rather than mean peer achievement, $\gamma_{\bar{y}}$ is identified (at least in a random assignment setting absent correlated unobserved shared inputs).

As discussed by Manski (1993) and others, endogenous and exogenous peer effects have quite different implications for policy, so addressing the simultaneity problem may be important. In particular, endogenous effects entail *social multipliers*, whereas exogenous effects do not. A social multiplier occurs when the improvement to one student's achievement leads to improvements in their peers. This would multiply any redistribution of resources among students. To consider an example, suppose there are only spillovers from peer ability. If all students benefit from higher peer ability, switching from tracked to mixed-ability classrooms leads to improvements for lower-ability students at the expense of higher-ability students. In contrast, if social multiplier effects exist, the losses (gains) to higher- (lower-) ability students from lower- (higher-) ability peers are partially offset by improvements (losses) in the achievement of their lower- (higher-) ability peers through the social multiplier.

Evidence of Social Effects

The vast majority of achievement peer effect studies focus on determining the social effect of peers. This reduced form estimate of peer effects is intuitively appealing, as many assignment policies are based directly on student observable Xs, as described in the above model. The actual mechanism of peer influence (endogenous or exogenous effects) may not be important. Similarly, whether the peer effect derives directly from the observable characteristic or correlated unobservable characteristics of the students may be a secondary concern. For instance, if the policy question is racial integration, it may not matter whether the

peer effect derives from race or other unobservables correlated with race.

Given that the objective is to determine the existence of a peer effect based on observable attributes of the students, selection of students into peer groups is the key challenge for these studies. While studies have found evidence of peer effects in the college setting using random assignment (See Carrell et al. 2008; Sacerdote 2001; Zimmerman 2003), random assignment to peer groups is much rarer in elementary and secondary school settings. Tennessee's Student Teacher Achievement Ratio project (Project STAR) is a notable exception. In this experiment, elementary students were randomly assigned to classrooms within schools. Graham (2008) and Boozer and Cacciola (2001) exploit the random assignment along with random variation in class size and find evidence of significant effects of being grouped with higher ability peers on student achievement.

Similar in spirit, quasi-experimental designs often exploit longitudinal data on student achievement to isolate plausibly random variation in peer groups to identify peer effects. For instance, Hoxby (2000) and Hanushek et al. (2009) attempt to isolate idiosyncratic variation in year-group-level peer composition across cohorts within schools to identify peer effects using Texas public school administrative data. The basic intuition is that while students may be systematically assigned to classrooms within a school, the year-group-level peer groups in a given year vary only for random reasons, such as differences in the birth cohort. In other settings, Lavy and Schlosser (2007) and Lavy et al. (2008) consider peer effects in Israeli schools; Ammermueller and Pischke (2009) provide interesting across country comparisons, relying on random assignment within schools.

While most studies focus on observable peer characteristics, Arcidiacono et al. (2010) develop an innovative, iterative approach to recovering spillovers from unobservable peer "ability", recovered as a persistent component of peer achievement (See Burke and Sass 2006, for an application in the elementary setting).

Behavioural Spillovers

While the above-mentioned studies provide important and compelling evidence of the existence of peer effects, they do not distinguish between spillovers deriving from peer characteristics and endogenous peer effects that would arise through behavioural spillovers among students within peer groups. Increasingly, evidence suggests that behavioural spillovers in the classroom exist.

Kinsler (2010) and Figlio (2007) show that having disruptive peers negatively affects student achievement. Lavy et al. (2008) find that having a higher proportion of repeaters has negative consequences on the classroom environment (as measured through survey questions). In related work, Lavy and Schlosser (2007) find that having more girls improves the classroom environment. Bishop et al. (2003) use unique survey data on high school students and find that peer pressure is a significant determinant of student effort and achievement at school. Fryer and Torelli (2005) find racial disparities in how academic achievement affects popularity, positing that black students face different social pressures than white students. Cipollone and Rosolia (2007), Gaviria and Raphael (2001), Nakajima (2007), Krauth (2005) find evidence that peers affect the decision to drop out of high school, and other behaviours, such as alcohol consumption, smoking or drug abuse, that may affect school achievement.

The evidence above strongly suggests that peer behaviour may affect students' behaviours in various ways. Absent direct measures of these behaviours, the achievement production framework captures these in peer achievement. For instance, harder-working, better-behaved students are higher-achieving. Students may benefit simply because the classroom learning environment with harder working (higherachieving) peers is more productive (e.g. Lazear 2001). Students could also benefit if they work harder because their peers are working harder, as implied by the other types of behavioural spillovers discussed above (See Bishop 2006; Akerlof and Kranton 2002).

Yet, because of the difficult simultaneity and the lack of clear exclusion restrictions, there is

little evidence regarding the effect of contemporaneous peer achievement on a student's achievement. Cooley (2010) illustrates potential sources of exclusion restrictions in the achievement production context. For instance, Cooley (2010) uses a policy that has differential effects on student incentives within the same classroom to address the simultaneity problem in achievement and finds evidence of large endogenous peer effects. This strategy is similar in spirit to strategies used in studies of behavioral peer effects in other settings, such as Cipollone and Rosolia (2007).

An alternative potential source of instruments is offered by partially overlapping peer groups, as discussed in Bramoulle et al. (2009), Giorgi et al. (2010), Cohen-Cole (2006) and others. The intuition (in its most basic form) is that if students A and B are grouped together in one setting, students B and C in another setting and A and C are never grouped together, then A provides an exclusion for determining the effect of B on C. This follows because A only affects C through his affect on B.

Nonlinearities

A well-known limitation of the linear-in-means model of achievement production with peer spillovers is that average achievement remains unchanged regardless of the grouping; the benefits to one group are perfectly offset by the losses to another. The literature recognizes this shortcoming and generally pays considerable attention to potential heterogeneity in peer effects. Studies find evidence of heterogeneity by race, gender and student ability (e.g. Hanushek et al. 2009; Gibbons and Telhaj 2006; Cooley 2010; Hoxby and Weingarth 2005; Lavy et al. 2008). In perhaps the most systematic investigation of nonlinearities in peer effects, Hoxby and Weingarth (2005) investigate a variety of potential models of peer effects and find that accurately accounting for nonlinearities in spillovers from lagged peer achievement (rather than just focusing on averages) eliminates much of the apparent spillovers from other observable peer characteristics.

Policy Implications

In contrast to the event studies described above, a limitation of the peer effects in achievement production studies is that generally the parameters estimated in these models are not directly applicable to reassignment policies. For instance, estimates of the effect of racial composition conditional on socioeconomic status may be interesting in principle, but not a feasible policy to implement in practice if race is correlated with SES. Furthermore, while studies generally find that achievement is negatively affected by higher percentages of nonwhite students (see, for instance, Hanushek et al. 2009), it would not be possible to lower the nonwhite percentage for all students. Put differently, these parameters can only be interpreted in a partial equilibrium sense.

In part, this limitation of the achievement production framework can be overcome by simulating the effect of different assignment policies by, for instance, randomly assigning students to classrooms and predicting the resulting achievement using the achievement production parameter estimates, as in Cooley (2010). Of course, the related challenge with moving to this type of general equilibrium context is that students cannot simply be sorted at will by policy makers. Thus, ultimately incorporating peer effect estimates into some of the existing general equilibrium locational sorting frameworks (e.g. Bayer et al. 2007; Ferreyra 2007; Nechyba 2000) is likely to be a useful way forward.

Graham et al. (2008) offer an innovative approach to this problem, moving beyond the achievement production framework to focus on estimating policyrelevant parameters. They directly characterize the effect of reallocations of students on the distribution of outcomes, particularly focusing on a model with two types of student (e.g. white and nonwhite). An appealing feature of this method is that it does not rely on distinguishing between the different micromechanisms of peer influence.

While this is a very useful approach, Cooley (2009) shows that when endogenous peer effects are heterogeneous (implying heterogeneous social multipliers) and there is sorting in the data, reduced

form estimates of the social effect of peers are often not sufficient to determine the effects of reallocations. Intuitively this follows because reassignment of students to classrooms also, by necessity, reallocates the unobserved group input (such as teacher quality) among students. This creates social multipliers that the reduced form estimator, even if estimated very flexibly, generally cannot predict when there is heterogeneity in student responses to peers. Evidence in the literature on exogenous peer effects and behavioural spillovers discussed above is strongly suggestive of heterogeneous endogenous peer effects. However, whether this is quantitatively important for understanding distributional consequences of reallocations remains to be determined.

Another important area of future research is likely to be the dynamics of peer effects, i.e. how peer effects vary with student age and over time. Vigdor and Nechyba (2007) and Carrell et al. (2008) find evidence that the effect of peers may persist over time. Studies, such as those by Todd and Wolpin (2003) and Cunha and Heckman (2007), show that the schooling dynamics generally are important for understanding human capital accumulation. The logic in these studies may extend to peer effects in important ways. Research on how the history of peer inputs determines achievement and the nature of social interactions in the classroom may further help inform school policy related to the age at which school choice policies, desegregation or academic tracking should be implemented.

In conclusion, it is worth noting that a better understanding of the role of peers in education production has implications far beyond the types of policies described above that are directly aimed at regrouping students. Many policies also indirectly affect student groupings. Most notably, various school choice mechanisms may have big effects on the composition of schools and classrooms. Evidence of peer effects also helps inform policies that are not directly targeted at schools. For instance, given that school composition is often closely tied to residential location, peer effects are important for understanding the broader implications of changes in property taxes or other policies that affect residential sorting patterns.

See Also

- ▶ Gender Differences (Experimental Evidence)
- ▶ School Choice and Competition

Bibliography

Akerlof, G.A., and R.E. Kranton. 2002. Identity and schooling: Some lessons for the economics of education. *Journal of Economic Literature* 40(4): 1167–1201.

Ammermueller, A., and J.-S. Pischke. 2009. Peer effects in European primary schools: Evidence from PIRLS. *Journal of Labor Economics* 27: 315–348.

Arcidiacono, P., Foster, G., Goodpaster, N., and Kinsler, J. 2010. Estimating spillovers using panel data. Working Paper.

Bayer, P., F. Ferreira, and R. McMillan. 2007. A unified framework for measuring preferences for schools and neighborhoods. *Journal of Political Economy* 115(4): 588–638.

Bishop, J. 2006. Chapter 15: Drinking from the fountain of knowledge: Student incentive to study and learn – externalities, information problems and peer pressure. In *Handbook of the economics of education*, vol. 2, ed. Eric A. Hanushek and Finis Welch, 909–944. St. Louis, MO: Elsevier.

Bishop, J.H., Bishop, M., Gelbwasser, L., Green, S., and Zuckerman, A. 2003. Nerds and freaks: A theory of student culture and norms. Brookings Papers on Education Policy.

Boozer, M.A., and Cacciola, S.E. 2001. Inside the black box of Project Star: Estimation of peer effects using experimental data. Yale Economic Growth Center Discussion Paper No. 832.

Bramoulle, Y., H. Djebbari, and B. Fortin. 2009. Identification of peer effects through social networks. *Journal of Econometrics* 150: 41–55.

Brock, W.A., and S.N. Durlauf. 2001a. Discrete choice with social interactions. *The Review of Economic Studies* 68(2): 235–260.

Brock, W.A., and S.N. Durlauf. 2001b. Interactions-based models. In *Handbook of econometrics*, vol. 5, ed. J. Heckman and E. Leamer, 3297–3380. Amsterdam: Elsevier.

Burke, M.A., and Sass, T.R. 2006. Classroom peer effects and student achievement. Working papers, Department of Economics, Florida State University.

Carrell, S.E., Fullerton, R.L., and West, J.E. 2008. Does your cohort matter? Measuring peer effects in college achievement. NBER Working Papers 14032, National Bureau of Economic Research, Inc.

Cipollone, P., and A. Rosolia. 2007. Social interactions in high school: Lessons from an earthquake. *American Economic Review* 97(3): 948–965.

Cohen-Cole, E. 2006. Multiple groups identification in the linear-in-means model. *Economics Letters* 92(2): 157–162.

C

Cooley, J.C. 2009. Can achievement peer effect estimates inform policy? A view from inside the black box. Working Paper.
Cooley, J.C. 2010. Desegregation and the achievement gap: Do diverse peers help? Working Paper.
Cunha, F., and J. Heckman. 2007. The technology of skill formation. *American Economic Review* 97(2): 31–47.
Ferreyra, M.M. 2007. Estimating the effects of private school vouchers in multidistrict economies. *American Economic Review* 97(3): 789–817.
Figlio, D.N. 2007. Boys named sue: Disruptive children and their peers. *Education Finance and Policy* 2(4): 376–394. http://www.mitpressjournals.org/doi/abs/10.1162/edfp.2007.2.4.376.
Fryer Jr., R.G., and Torelli, P. 2005. An empirical analysis of 'acting white'. NBER Working Paper No. 11334.
Gamoran, A. 2009. Tracking and inequality: New directions for research and practice. WCER Working Paper No. 2009-6.
Gaviria, A., and S. Raphael. 2001. School-based peer effects and juvenile behavior. *The Review of Economics and Statistics* 83(2): 257–268.
Gibbons, S., and Telhaj, S. 2006. Peer effects and pupil attainment: Evidence from secondary school transition. London School of Economics. CEP Working Papter.
Giorgi, G.D., M. Pellizzari, and S. Redaelli. 2010. Identification of social interactions through partially overlapping peer groups. *American Economic Journal: Applied Economics* 2(2): 241–275.
Graham, B.S. 2008. Identifying social interactions through conditional variance restrictions. *Econometrica* 76(3): 643–660.
Graham, B.S., Imbens, G., and Ridder, G. 2008. Measuring the average outcome and inequality effects of segregation in the presence of social spillovers. Working Paper.
Guryan, J. 2004. Desegregation and black dropout rates. *American Economic Review* 94(4): 919–943.
Hanushek, E., J. Kain, and S. Rivkin. 2009. New evidence about brown v. board of education: The complex effects of school racial composition on achievement. *Journal of Labor Economics* 27(3): 349–383.
Harris, D. 2007. Chapter 31: Educational outcomes of disadvantaged students: From desegregation to accountability. In *AEFA handbook of research in education finance and policy*, ed. H. Ladd and E. Fiske. Hillsdale, NJ: Laurence Erlbaum.
Hoxby, C. 2000. Peer effects in the classroom: Learning from gender and race variation. Working Paper 7867, National Bureau of Economic Research.
Hoxby, C.M., and Weingarth, G. 2005. Taking race out of the equation: School reassignment and the structure of peer effects. Working Paper.
Kinsler, J. 2010. School discipline: A source or salve for the racial achievement gap? Working Paper.
Krauth, B.V. 2005. Peer effects and selection effects on smoking among canadian youth. *The Canadian Journal of Economics/Revue canadienne d'Economique* 38(3): 735–757.
Lavy, V., and Schlosser, A. 2007. Mechanisms and impacts of gender peer effects at school. NBER Working Papers 13292, National Bureau of Economic Research, Inc.
Lavy, V., Paserman, M.D., and Schlosser, A. 2008. Inside the black box of ability peer effects: Evidence from variation in low achievers in the classroom. NBER Working Papers 14415, National Bureau of Economic Research, Inc.
Lazear, E.P. 2001. Educational production. *Quarterly Journal of Economics* 116(3): 777–803.
Manski, C. 1993. Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies* 60(3): 531–542.
Moffitt, R.A. 2001. Policy interventions, low-level equilibria and social interactions. In *Social dynamics*, ed. S.N. Durlauf and H.P. Young, 45–82. Washington, DC: Brookings Institution.
Nakajima, R. 2007. Measuring peer effects on youth smoking behaviour. *Review of Economic Studies* 74(3): 897–935.
Nechyba, T.J. 2000. Mobility, targeting, and private-school vouchers. *American Economic Review* 90(1): 130–146.
Sacerdote, B.I. 2001. Peer effects with random assignment: Results for Dartmouth room-mates. *Quarterly Journal of Economics* 116: 681–704.
Todd, P.E., and K.I. Wolpin. 2003. On the specification and estimation of the production function for cognitive achievement. *Economic Journal* 113(485): F3–F33.
Vigdor, J., and T. Nechyba. 2007. Peer effects in North Carolina Public Schools. In *Schools and the equal opportunity problem*, ed. P. Peterson and L. Woessmann. Cambridge, MA: MIT Press.
Zimmerman, D.J. 2003. Peer effects in academic outcomes: Evidence from a natural experiment. *Review of Economics and Statistics* 85(1): 9–23.

Cliffe Leslie, Thomas Edward (1827–1882)

J. Maloney

T.E. Cliffe Leslie was both the pioneer (with J.K. Ingram) and the most radical member of the English Historical School. Born in Co. Wexford, he was educated at King William's College, Isle of Man, and at Trinity College, Dublin. In 1853 he became professor of jurisprudence and political economy at Queen's College, Belfast. His inaugural lecture 'The Military Systems of Europe Economically Considered' was published in

1856 and set the empirical, comparative tone that informed all his work. It was, however, Leslie's Irish context that did most to sharpen his onslaught on orthodox economics. To the Irish tenant-farmers, lacking either security of tenure or the right to be compensated for improvements they had made, liberal economists offered only free trade and the assurance that no good could come from specific legislation for Ireland. Thus Robert Lowe (shortly to become Gladstone's Chancellor) in 1868 urged Parliament to oppose land reform 'with the principles of political economy'. That Mill dissented, supporting what eventually became the Irish Land Act of 1870, was perhaps the crucial episode in Leslie's becoming a self-proclaimed disciple of Mill.

In 1870 Leslie published a volume of essays entitled *Land Systems and the Industrial Economy of Ireland, England and Continental Countries*. Highly praised by Mill in the *Fortnightly Review*, it was the last and most important of his works directly on the Irish question. In the same year he fired his opening salvo in the English *Methodenstreit* with 'The Political Economy of Adam Smith' (1888). Here Leslie lauded Smith as an inductive, historically minded economist whose brand of economics should never have been supplanted by the grotesque abstractions and unbalanced methodology of Ricardo. In 'On the Philosophical Method of Political Economy' (1888), Leslie reiterated that man was not, as the classical economists had assumed, a being whose 'great variety of different and heterogeneous motives' could be compounded into a homogenous desire for wealth when deductive analysis was required. He went on to attack orthodox political economy for its failure – inability, indeed, as it stood constituted – to go behind the individual economic agent and weigh up the social and historical forces moulding his actions and preferences. But perhaps his most radical paper, and certainly the one which strikes the strongest chord today, is 'The Known and the Unknown in Economics' (1888), in which the limits of the economist's information – and the blithe unconcern of many economists about the fact that these limits exist – are laid sharply bare. Here, again, historical relativism is the mainspring, Leslie arguing that only in a primitive village economy, with predictable 'reproduction' of economic activities and prices set by custom, can the economist have something approaching complete knowledge. When prices are set in a competitive market, when credit and default weaken certainty and trust, when technical change accelerates to the point where products rapidly become obsolete, then universalist economic 'laws' represent little more than an (unwitting) confession of ignorance as to the really important features of any particular episode until it can be written up from a historian's perspective.

Leslie's intended *magnum opus*, a work on the economic and legal history of England, had only reached manuscript stage when he lost it in France in 1872. For the remaining ten years of his life, his contribution was more critical than constructive. However, in the 1870s, radical criticism of deductive economics was badly needed, as both Marshall and John Neville Keynes were later to admit.

Selected Works

1888. *Essays in political economy.* Dublin: Hodges, Figgis & Co.

References

Hutchison, T.W. 1978. *On revolutions and progress in economic knowledge*. Cambridge: Cambridge University Press.

Koot, G. 1975. T.E. Cliffe Leslie, Irish social reform, and the origins of the English historical school of economics. *History of Political Economy* 7(3): 312–336.

Climate Change, Economics of

Lawrence H. Goulder and William A. Pizer

Abstract

Climate-change economics attends to the various threats posed by global climate change by offering theoretical and empirical insights relevant to the design of policies to reduce, avoid,

or adapt to such change. This economic analysis has yielded new estimates of mitigation benefits, improved assessments of policy costs in the presence of various market distortions or imperfections, better tools for making policy choices under uncertainty, and alternative mechanisms for allowing flexibility in policy responses. These contributions have influenced the formulation and implementation of a range of climate-change policies at domestic and international levels.

Keywords

Carbon emissions tax; Climate change, economics of; Computable general equilibrium (CGE) models; Contingent valuation; Discount rate; Global warming; Hedonic approach; Integrated assessment models; Intergenerational equity; Learning-by-doing; Monte Carlo methods; Price-based vs. quantity-based policies; Production function approach; Technology policy; Time preference; Tradable emission permits; Uncertainty

JEL Classification

Q54

The prospect of global climate change has emerged as a major scientific and public policy issue. Scientific studies indicate that human-caused increases in atmospheric concentrations of carbon dioxide (largely from fossil-fuel burning) and of other greenhouse gases are leading to warmer global surface temperatures. Possible current-century consequences of this temperature increase include increased frequency of extreme temperature events (such as heat waves), heightened storm intensity, altered precipitation patterns, sea-level rise, and reversal of ocean currents. These changes, in turn, can have significant impacts on the functioning of ecosystems, the viability of wildlife and the well-being of humans.

There is considerable disagreement within and among nations as to what policies, if any, should be introduced to mitigate and perhaps prevent climate change and its various impacts. Despite the disagreements, in recent years we have witnessed the gradual emergence of a range of international and domestic climate- change policies, including emission-trading programmes, emission taxes, performance standards, and technology-promoting programmes.

Beginning with William Nordhaus's 'How fast should we graze the global commons?' (Nordhaus 1982), climate-change economics has focused on diagnosing the economic underpinnings of climate change and offering positive and normative analyses of policies to confront the problem. While overlapping with other areas of environmental economics, it has a unique focus because of distinctive features of the climate problem – including the long time-scale, the extent and nature of uncertainties, the international scope of the issue, and the uneven distribution of policy benefits and costs across space and time.

In our discussion of the economics of climate change, we begin with a brief account of alternative economic approaches to measuring the benefits and costs associated with reducing greenhouse gas emissions, followed by a discussion of uncertainties and their consequences. We then present issues related to policy design, including instrument choice, flexibility, and international coordination. The final section offers general conclusions.

Assessing the Benefits and Costs of Climate Change Mitigation

Climate Change Damages and Mitigation Benefits

As noted, the potential consequences of climate change include increased average temperatures, greater frequency of extreme temperature events, altered precipitation patterns, and sea-level rise. These biophysical changes affect human welfare. While the distinction is imperfect, economists divide the (often negative) welfare impacts into two main categories: *market* and *non-market* damages.

Market Damages As the name suggests, market damages are the welfare impacts stemming from

changes in prices or quantities of marketed goods. Changes in productivity typically underlie these impacts. Often researchers have employed climate-dependent production functions to model these changes, specifying wheat production, for example, as a function of climate variables such as temperature and precipitation. In addition to agriculture, this approach has been applied in other industries including forestry, energy services, water utilities and coastal flooding from sea-level rise (see, for example, Smith and Tirpak 1989; Yohe et al. 1996; Mansur et al. 2005).

The production function approach tends to ignore possibilities for substitution across products, which motivates an alternative, hedonic approach (see, for example, Mendelsohn et al. 1994; Schlenker et al. 2005). Applied to agriculture, the hedonic approach aims to embrace a wider range of substitution options, employing cross-section data to examine how geographical, physical, and climate variables are related to the prices of agricultural land. On the assumption that crops are chosen to maximize rents, that rents reflect the productivity of a given plot of land relative to that of marginal land, and that land prices are the present value of land rents, the impact of climate variables on land prices is an indicator of their impact on productivity after crop-substitution is allowed for.

Non-market Damages Non-market damages include the direct utility loss stemming from a less hospitable climate, as well as welfare costs attributable to lost ecosystem services or lost biodiversity. For these damages, revealed-preference methods face major challenges because non-market impacts may not leave a 'behavioural trail' of induced changes in prices or quantities that can be used to determine welfare changes. The loss of biodiversity, for example, does not have any obvious connection with price changes or observable demands. Partly because of the difficulties of revealed-preference approaches in this context, researchers often employ stated-preference or interview techniques – most notably the contingent valuation method – to assess the willingness to pay to avoid non-market damages (see, for example, Smith 2004).

Cost Assessment

The costs of avoiding emissions of carbon dioxide, the principal greenhouse gas, depend on substitution possibilities on several margins: the ability to substitute across different fuels (which release different amounts of carbon dioxide per unit of energy), to substitute away from energy in general in production, and to shift away from energy-intensive goods. The greater the potential for substitution, the lower are the costs of meeting a given emission-reduction target.

Applied models have taken two main approaches to assessing substitution options and costs. One approach employs 'bottom-up' energy technology models with considerable detail on the technologies of specific energy processes or products (for example, Barretto and Kypreos 2004). The models tend to concentrate on one sector or a small group of sectors, and offer less information on abilities to substitute from energy in general or on how changes in the prices of energy-intensive goods affect intermediate and final demands for those goods.

The other approach employs 'top down' economy-wide models, which include, but are not limited to, computable general equilibrium (CGE) models (see, for example, Jorgenson and Wilcoxen 1996; Conrad 2002). An attraction of these models is their ability to trace relationships between fuel costs, production methods, and consumer choices throughout the economy in an internally consistent way. However, they tend to include much less detail on specific energy processes or products. Substitution across fuels is generally captured through smooth production functions rather than through explicit attention to alternative discrete processes. In recent years, attempts have been made to reduce the gap between the two types of models. Bottom-up models have gained scope, and top-down models have incorporated greater detail (see, for example, McFarland et al. 2004).

Because climate depends on the atmospheric stock of greenhouse gases, and because for most gases the residence times in the atmosphere are hundreds (and in some cases, thousands) of years,

climate change is an inherently long-term problem and assumptions about technological change are particularly important. The modelling of technological change has advanced significantly beyond the early tradition that treated technological change as exogenous. Several recent models allow the rate or direction of technological progress to respond endogenously to policy interventions. Some models focus on *R&D-based* technological change, incorporating connections between policy interventions, incentives to research and development, and advances in knowledge (see, for example, Goulder and Schneider 1999; Nordhaus 2002; Buonanno et al. 2003; Popp 2004). Others emphasize *learning-by-doing-based* technological change where production cost falls with cumulative output, in keeping with the idea that cumulative output is associated with learning (for example, Manne and Richels 2004). Allowing for policy-induced technological change tends to yield lower (and sometimes significantly lower) assessments of the costs of reaching given emission-reduction targets than do models in which technological change is exogenous.

Integrated Assessment

While the cost models described above are useful for evaluating the cost- effectiveness of alternative policies to achieve a given emissions target, the desire to relate costs to mitigation benefits (avoided damages) has spawned the development of *integrated assessment models.* These models link greenhouse gas emissions, greenhouse gas concentrations, and changes in temperature or precipitation, and they consider how these changes feed back on production and utility. Many of the integrated assessment models are optimization models that solve for the emissions time-path that maximizes net benefits, in some cases under constraints on temperature or concentration (see, for example, Nordhaus 1994).

Dealing with Uncertainty

The uncertainties about both the costs and the benefits from reduced climate change are vast. In a recent meta-analysis examining 28 studies' estimated benefits from reduced climate change (Tol 2005), the 90% confidence interval for the benefit estimates ranged from − $10 to + $350 per ton of carbon, with a mode of $1.50 per ton. On the cost side, a separate study found marginal costs of between $10 and $212 per ton of carbon for a ten per cent reduction in 2010 (Weyant and Hill 1999).

Uncertainty and the Stringency of Climate Policy

Increasingly sophisticated numerical models have attempted to deal explicitly with these substantial uncertainties regarding costs and benefits. Some provide an uncertainty analysis using Monte Carlo simulation, in which the model is solved repeatedly, each time using a different set of parameter values that are randomly drawn from pre-assigned probability distributions. This approach produces a probability distribution for policy outcomes that sheds light on appropriate policy design in the face of uncertainty. Other models incorporate uncertainty more directly by explicitly optimizing over uncertain outcomes. These models typically call for a more aggressive climate policy than would emerge from a deterministic analysis. Nordhaus (1994) employs an integrated climate-economy model to compare the optimal carbon tax in a framework with uncertain parameter values with the optimal tax when parameters are set at their central values. In this application, an uncertainty premium arises: the optimal tax is more than twice as high in the former case as in the latter, and the optimal amount of abatement is correspondingly much greater. The higher optimal tax could in principle be due to uncertainty about any parameter whose relationship with damages is convex, thus yielding large downside risks relative to upside risks. In the Nordhaus model, the higher optimal tax stems primarily from uncertainty about the discount rate (Pizer 1999).

The Choice of Discount Rate Under Uncertainty

The importance of the discount rate arises because greenhouse gases persist in the atmosphere for a century or more, and therefore mitigation benefits

must be measured on dramatically different timescales from those of ordinary environmental problems. A prescriptive approach links the discount rate to subjective judgements about intergenerational equity as indicated by a pure social rate of time preference (see, for example, Arrow et al. 1996). A descriptive approach relates the discount rate to future market interest rates. Under both approaches, significant uncertainties surround the discount rates. Recent work by Weitzman (1998) points out that a rate lower than the expected value should be employed in the presence of such uncertainty, a reflection of the relationships among the discount *factor*, the discount *rate*, and the time interval over which discounting applies. Put simply, the discount factor e^{-rt} is an increasingly convex function of the interest rate r as the period of discounting t increases. This implies that in the presence of uncertainty the certainty- equivalent discount rate is lower than the expected value of the discount rate: that is, $ln(\mathrm{E}[e^{-rt}])/t < \mathrm{E}[r]$. The difference between the appropriate, certainty-equivalent rate and the expected value of the discount rate widens the longer the time horizon is. While Weitzman focuses on a single uncertain rate, Newell and Pizer (2003a) show that, under reasonable specifications of uncertainty about the evolution of future market rates, this approach doubles the expected marginal benefits from future climate change mitigation compared with the estimated benefits from an analysis that uses only the current rate.

Act Today or Wait for Better Information?

In addition to concerns about convexity and valuation, uncertainty raises important questions about whether and how much to embark on mitigation activities now as opposed to waiting until at least some uncertainty is resolved. Economic theory suggests that, in the absence of fixed costs and irreversibilities, society should mitigate (today) to the point where expected marginal costs and benefits are equal. Yet climate change inherently involves fixed costs and irreversible decisions both on the cost side, in terms of investments in carbon-free technologies, and on the benefit side, in terms of accumulated emissions. These features can lead to more intensive action or to inaction, depending on the magnitude of their respective sunk values (Pindyck 2000). Despite the ambiguous theory, empirically calibrated analytical and numerical models tend to recommend initiating reductions in emissions in the present, reflecting initially negligible marginal cost and non-negligible environmental benefits (Manne and Richels 2004; Kolstad 1996).

The Choice of Instrument for Climate-Change Policy

Policymakers can consider a range of potential instruments for promoting reductions in emissions of greenhouse gases. Alternatives include emissions taxes, abatement subsidies, emission quotas, tradable emission allowances, and performance standards. Policymakers also can choose whether to apply a given instrument to emissions directly (as with an emission-trading programme) or instead to pollution-related goods or services (as with a fuel tax or technology subsidy).

Initial economic analyses of climate-change policy tended to focus on a carbon tax because it was relatively easy to model and implement. This is a tax on fossil fuels – oil, coal, and natural gas – in proportion to their carbon content. Because combustion of fossil fuels or their refined fuel products leads to carbon dioxide (CO_2) emissions proportional to carbon content, a carbon tax is effectively a tax on CO_2 emissions. In the simplest analysis, a carbon tax set equal to the marginal climate-related damage from carbon combustion would be efficiency-maximizing. However, in more complex analyses – where additional dimensions such as uncertainty, other market failures, and distributional impacts are taken into account – the superiority of such a carbon tax is no longer assured. We now consider these other dimensions and their implications for instrument choice.

Prices (Taxes) vs. Quantities (Tradable Allowances) in the Presence of Uncertainty

Theoretical and empirical work by Kolstad (1996) and Newell and Pizer (2003b) suggests that the

marginal benefit (avoided damage) schedule for emissions reductions is relatively flat. Weitzman's (1974) seminal analysis indicates that under these circumstances expected welfare losses are smaller from a price-based instrument like a carbon tax than from a quantity-based instrument like emission quotas or a system of tradable emission allowances. That is, it is preferable to let levels of emissions remain uncertain (which is the result under a tax) than to let the marginal price of emission reductions remain uncertain (which is the result under a quota). Despite these economic welfare arguments, and recent work on hybrid approaches (Pizer 2002), many environmental advocates prefer the quantity-based approach precisely because it removes uncertainty about the level of emissions.

Fiscal Impacts and Instrument Choice

A second issue stems from the potential for policies such as carbon taxes and auctioned permits to generate revenues. A number of studies show that using such revenues to finance reductions in pre-existing distortionary taxes on income, sales, or payroll can achieve given environmental targets at lower cost – perhaps substantially lower cost – than other policies (see, for example, Goulder et al. 1999; Parry et al. 1999; Parry and Oates 2000). Therefore, carbon taxes and auctioned permit programmes that employ their revenues this way will lower the excess burden from prior taxes, giving them a significant cost-advantage. Correspondingly, subsidies to emission reductions or to new, 'clean' technologies will have a cost disadvantage associated with the need to raise distortionary taxes to finance these policies.

Distributional Considerations

Despite these attractions of revenue-raising policies such as carbon taxes and auctioned tradable allowance systems, trading programmes with freely distributed permits have achieved greater popularity among policymakers. In New Zealand, for example, industry opposition led the government to drop its proposed carbon tax in 2005. At the same time, the European Union has, and Canada is planning, trading programmes where tradable permits are freely distributed, in line with virtually all conventional pollution trading programmes in the United States.

The politics may reflect differences between systems of freely allocated allowances and systems with auctioned allowances (or carbon taxes) in terms of the distribution of the regulatory burden. Under both types of emission-permit system, profit-maximizing firms will find it in their interest to raise output prices based on the new, non-zero cost associated with carbon emissions. If the allowances are given out free, firms can retain rents associated with the higher output prices, and this offsets other compliance costs. In contrast, if the allowances are auctioned, firms do not capture these rents. Thus, firms bear a considerably smaller share of the regulatory burden in the case of freely allocated permits. Indeed, Bovenberg and Goulder (2001) show that freely allocating all carbon permits to US fossil fuel suppliers generally will cause those firms to enjoy *higher* profits than in the absence of a permit system; and freely allocating less than a fifth of the permits may be sufficient to keep profits from falling. These considerations reveal a potential trade-off between efficiency and political feasibility: the revenue-raising policies (taxes and auctioned permits) are the most cost-effective, while the non-revenue-raising policies (freely distributed permits) have distributional consequences that may reduce political resistance.

Emissions Instruments vs. Technology Instruments

As noted in the cost discussion, the long-term nature of the climate-change problem makes technological change a central issue in policy considerations. Economic analysis suggests that both 'direct emissions policies' and 'technology-push policies' are justified on efficiency grounds to correct two distinct market failures. Direct emissions policies (emission trading or taxes) gain support from the fact that combustion of fossil fuels and other greenhouse-gas-producing activities generate negative externalities in the form of climate change-related damages. Technology-push policies (technology and R&D incentives)

gain support from the fact that not all of the social benefits from the invention of a new technology can be appropriated by the inventor. The latter argument applies to research and development more generally, and is especially salient if the first market failure is not fully corrected (Fischer 2004a). Numerical assessments reveal substantial cost-savings from combining the two types of policy (Fischer and Newell 2005; Schneider and Goulder 1997).

Policy Designs to Enhance Flexibility

The previous discussion indicates that no single instrument is best along all important policy dimensions, including cost uncertainty, fiscal interactions, distribution and technology development. A further issue in policy choice is how to give regulated firms or nations the flexibility to seek out mitigation opportunities wherever and whenever they are cheapest. For both price- and quantity-based policies, flexibility is enhanced through broad coverage: specifically, by including in the programme as many emissions sources as possible and by providing opportunities for regulated sources to offset their obligations through relevant activities outside the programme. For quantity-based programmes, flexibility can also be promoted through provisions allowing trading of allowances across gases, time, and national boundaries. Such flexibility is automatically provided by price-based programmes simply because they involve no quantitative emissions limits. Importantly, as quantity-based programmes provide these additional dimensions of flexibility, they reduce the efficiency arguments for price-based policies in the face of uncertainty voiced in the preceding section by providing opportunities to adjust to idiosyncratic cost shocks across time, space and industry (Jacoby and Ellerman 2004).

Flexibility Over Gases and Sequestration

So far we have focused almost exclusively on emissions of carbon dioxide from the burning of fossil fuels as both the cause of human-induced climate change and the object of any mitigation policy. Yet emissions of a number of other gases (as well as non-energy-related emissions of carbon dioxide) contribute to the problem and possibly the solution, particularly in the short run. Models suggest that half of the reductions achievable at costs of $5–$10 per ton of carbon dioxide equivalent arise from gases other than carbon dioxide. In addition, carbon sequestration can be part of the solution. Biological sequestration (for example, through afforestation) has been cited as a particularly inexpensive response to climate change (Sedjo 1995; Richards and Stavins 2005). Geological sequestration (for example, injection into depleted oil or gas reservoirs) represents a very expensive proposition now, but could be an important component of a long-term policy solution if costs decline (Newell and Anderson 2004).

Four issues can complicate the inclusion of these activities: monitoring, baselines, comparability and, in some cases, liability. First, some of these sources are fugitive emissions that are difficult to monitor at any point in the product cycle. Second, some activities, especially those involving fugitive emissions, are often left unregulated but allowed to enter as ‘offsets’, requiring a counterfactual baseline against which actual emissions levels can be measured. Fischer (2004b) evaluates various approaches to defining project baselines. Third, a problem of comparability arises with non-CO_2 gases because it is necessary to determine relative prices among greenhouse gases in a market-based programme. As a theoretical matter, the ratio of prices of a ton of current emissions of two different gases should be the ratio of the present value of damages from these emissions (Schmalensee 1993). In practice it is difficult to apply this formula because it requires a great deal of information about the damages and because it calls for time-varying trading ratios (Reilly et al. 2001), which implies significant administrative burdens. Under the Kyoto Protocol and the EU Emissions Trading Scheme, one set of trading ratios is used at all times, and the ratios are calculated by determining the ratio of warming impacts over a 100-year horizon beginning with the present time. Finally, a liability issue arises with regard to sequestration. For both biologically

and geologically sequestered carbon, a key question is who should be held liable for carbon dioxide that is released accidentally or otherwise.

Flexibility Over Time

While price policies naturally allow emissions to rise and fall in response to shocks over time, quantity-based policies must explicitly address the question of whether regulated sources can bank unused allowances for future use or, in some cases, borrow them from future allocations. In the climate change context, merely shifting emissions across time, as opposed to allowing accumulated emissions to vary, holds the environment harmless because climate consequences are generally due to accumulated concentrations, not annual emissions (Roughgarden and Schneider 1999, discuss the possibility of dependence on both accumulated concentrations and the rate of accumulation.) Such shifts across time might reflect either a more efficient choice of timing in response to capital turnover and technological progress (Wigley et al. 1996), or an attempt to ameliorate cost shocks (Williams 2002; Jacoby and Ellerman 2004). The rate of exchange between present and future emissions allowances need not be unity: Kling and Rubin (1997) show that the optimal rate at which banked allowances translate across periods should reflect the expected trend in marginal mitigation benefits, the interest rate, and decay rate of the accumulated gas.

Flexibility Over Location

The defining feature of the climate-change problem may be its intrinsically global nature. Greenhouse gases tend to disperse themselves uniformly around the globe. As a result, the climate consequences of a ton of emissions of a given greenhouse gas do not depend on the location of the source, either within or across national borders, and shifts in emissions across locations do not change global climate impacts. Under these circumstances, economic efficiency calls for making market- based systems as geographically broad as possible. It supports federal over regional policies, and international coordination over idiosyncratic domestic responses.

International Policy Initiatives and Coordination

International coordination is both crucial and exceptionally difficult to achieve. Studies indicate that the economic and social impacts of climate change would be distributed very unevenly across the globe, with the prospect of large damages to several nations in the tropics coupled with the potential for *benefits* to some countries in the temperate zones (see, for example, Tol 2005; Mendelsohn 2003). This uneven distribution makes achieving international coordination especially difficult.

The Kyoto Protocol is the first significant international effort to reduce greenhouse gas emissions. It assigns emission limits to participating industrialized countries for 2008–2012, but offers flexibility in allowing these countries to alter their limits by buying or selling emission allowances from other industrialized countries or by investing in projects that lead to emission reductions in developing countries. The importance of these flexibility mechanisms for dramatically lowering compliance costs in this international setting is well documented (Weyant and Hill 1999).

The Protocol has been criticized on the grounds that it imposes overly stringent emission-reduction targets and lacks a longer-term vision for action. In addition, a core feature of the Protocol – legally-binding emission limits – has been challenged on the grounds that such limits are not self-enforcing, an arguably necessary attribute in a world of sovereign nations (Barrett 2003). Some argue that the Protocol's project-based mechanisms for encouraging (but not requiring) emission reductions in developing countries are highly bureaucratic and cumbersome, consistent with our earlier comments about project-based programmes more generally. These criticisms have led to considerable research considering the Kyoto structure and comparing it with various alternative international approaches. Aldy et al. (2003) summarize more than a dozen alternatives, which include an international carbon tax and international technology standards.

A further major criticism is that the Protocol imposes no mandatory emissions limits on

developing countries, which collectively are expected to match industrialized countries in emissions of greenhouse gases by 2035. The desire to promote greater participation by developing countries, as well as to involve the United States in the international effort, has motivated considerable research examining, within a game-theoretic framework, the requirements for broader participation and for stable international coalitions (see, for example, Carraro 2003; Hoel and Schneider 1997; Tulkens 1998).

Conclusions

Climate-change economics has produced new methods for evaluating environmental benefits, for determining costs in the presence of various market distortions or imperfections, for making policy choices under uncertainty, and for allowing flexibility in policy responses. Although major uncertainties remain, it has helped generate important guidelines for policy choice that remain valid under a wide range of potential empirical conditions. It has also helped focus empirical work by making clear where better information about key parameters would be most valuable.

Clearly, many theoretical and empirical questions remain unanswered. We suggest (with some subjectivity) that there is a particularly strong need for advances in the integration of emissions policy and technology policy, in defining baselines that determine the extent of offset activities outside a regulated system, and in fostering international cooperation.

From 2003 until 2030 the world is poised to invest an estimated $16 trillion in energy infrastructure, with annual carbon dioxide emissions estimated to rise by 60%. How well economists answer important remaining questions about climate change could have a profound impact on the nature and consequences of that investment.

See Also

- ▶ Coalitions
- ▶ Computation of General Equilibria
- ▶ Contingent Valuation
- ▶ Diffusion of Technology
- ▶ Environmental Economics
- ▶ Energy Economics
- ▶ Hedonic Prices
- ▶ Learning-by-Doing
- ▶ Options
- ▶ Pigouvian Taxes
- ▶ Second Best
- ▶ Social Discount Rate
- ▶ Uncertainty

Acknowledgment The authors gratefully acknowledge very helpful comments on earlier drafts by Kenneth Arrow, Steven Durlauf, Raymond Kopp, Richard Morgenstern, Robert Stavins and Roberton Williams III.

Bibliography

Aldy, J., S. Barrett, and R. Stavins. 2003. Thirteen plus one: A comparison of alternative climate policy architectures. *Climate Policy* 3: 373–397.

Arrow, K., W. Cline, K.-G. Maler, M. Munasinghe, R. Squitieri, and J. Stiglitz. 1996. Intertemporal equity, discounting and economic efficiency. In *Climate change 1995 – Economic and social dimensions of climate change*, ed. J. Bruce, H. Lee, and E. Haites. Cambridge: Cambridge University Press.

Barrett, S. 2003. *Environment and statecraft*. New York: Oxford University Press.

Barretto, L., and S. Kypreos. 2004. Emissions trading and technology deployment in an energy-system 'bottom-up' model with technological learning. *European Journal of Operations Research* 158: 243–261.

Bovenberg, A., and L. Goulder. 2001. Neutralizing the adverse industry impacts of CO_2 abatement policies: What does it cost? In *Behavioral and distributional effects of environmental policies*, ed. C. Carraro and G. Metcalf. Chicago: University of Chicago Press.

Buonanno, P., C. Carraro, and E. Galeotti. 2003. Endogenous induced technical change and the costs of Kyoto. *Resource and Energy Economics* 25: 11–34.

Carraro, C. (ed.). 2003. *The endogenous formation of economic coalitions*. Northampton: Edward Elgar.

Conrad, K. 2002. Computable general equilibrium models in environmental and resource economics. In *The international yearbook of environmental and resource economics 2002–2003*, ed. T. Tietenberg and H. Folmer. Cheltenham: Edward Elgar.

Fischer, C. 2004a. *Emission pricing, spillovers, and public investment in environmentally friendly technologies*. Discussion Paper 04–02. Washington, DC: Resources for the Future.

Fischer, C. 2004b. Project-based mechanisms for emissions reductions: balancing trade-offs with baselines. *Energy Policy* 33:1807–1823.

Fischer, C., and R. Newell. 2005. *Environmental and technology policies for climate mitigation*. Working Paper. Washington, DC: Resources for the Future.

Goulder, L., and S. Schneider. 1999. Induced technological change and the attractiveness of CO_2 emissions abatement policies. *Resource and Energy Economics* 21: 211–253.

Goulder, L., I. Parry, R. Williams III, and D. Burtraw. 1999. The cost-effectiveness of alternative instruments for environmental protection in a second-best setting. *Journal of Public Economics* 72: 329–360.

Hoel, M., and K. Schneider. 1997. Incentives to participate in an international environmental agreement. *Environment and Resource Economics* 9: 153–170.

Jacoby, H., and A. Ellerman. 2004. The safety valve and climate policy. *Energy Policy* 32: 481–491.

Jorgenson, D., and P. Wilcoxen. 1996. Reducing U.S. Carbon emissions: An econometric general equilibrium assessment. In *Reducing global carbon dioxide emissions: Costs and policy options*, ed. D. Gaskins and J. Weyant. Stanford: Energy Modeling Forum/Stanford University.

Kling, C., and J. Rubin. 1997. Bankable permits for the control of environmental pollution. *Journal of Public Economics* 64: 101–115.

Kolstad, C. 1996. Learning and stock effects in environmental regulation: The case of greenhouse gas emissions. *Journal of Environmental Economics and Management* 31: 1–18.

Manne, A., and R. Richels. 2004. The impacts of learning-by-doing on the timing and costs of CO_2 abatement. *Energy Economics* 26: 603–619.

Mansur, E., R. Mendelsohn, and W. Morrison. 2005. *A discrete-continuous choice model of climate change impacts on energy*. New Haven: Yale School of Forestry and Environmental Studies.

McFarland, J., J. Reilly, and H. Herzog. 2004. Representing energy technologies in top-down economic models using bottom-up information. *Energy Economics* 26: 685–707.

Mendelsohn, R. 2003. Assessing the market damages from climate change. In *Global climate change: The science, economics, and politics*, ed. J. Griffin. Cheltenham: Edward Elgar.

Mendelsohn, R., W. Nordhaus, and D. Shaw. 1994. The impact of global warming on agriculture: A Ricardian analysis. *American Economic Review* 84: 753–771.

Newell, R., and W. Pizer. 2003a. Discounting the distant future: How much do uncertain rates increase valuations? *Journal of Environmental Economics and Management* 46: 52–71.

Newell, R., and W. Pizer. 2003b. Regulating stock externalities under uncertainty. *Journal of Environmental Economics and Management* 45: 416–432.

Newell, R., and S. Anderson. 2004. Prospects for carbon capture and storage technology. *Annual Review of Environment and Resources* 29: 109–142.

Nordhaus, W. 1982. How fast should we graze the global commons? *American Economic Review* 72: 242–246.

Nordhaus, W. 1994. *Managing the global commons*. Cambridge, MA: MIT Press.

Nordhaus, W. 2002. Modeling induced innovation in climate-change policy. In *Technological change and the environment*, ed. A. Grübler, N. Nakicenovic, and W.D. Nordhaus. Washington, DC: Resources for the Future.

Parry, I., and W. Oates. 2000. Policy analysis in the presence of distorting taxes. *Journal of Policy Analysis and Management* 19: 603–614.

Parry, I., R. Williams III, and L. Goulder. 1999. When can carbon abatement policies increase welfare? The fundamental role of distorted factor markets. *Journal of Environmental Economics and Management* 37: 52–84.

Pindyck, R. 2000. Irreversiblities and the timing of environmental policy. *Resource and Energy Economics* 22: 233–259.

Pizer, W. 1999. Optimal choice of policy instrument and stringency under uncertainty: The case of climate change. *Resource and Energy Economics* 21: 255–287.

Pizer, W. 2002. Combining price and quantity controls to mitigate global climate change. *Journal of Public Economics* 85: 409–434.

Popp, D. 2004. ENTICE: Endogenous technological change in the DICE model of global warming. *Journal of Environmental Economics and Management* 48: 742–768.

Reilly, J., M. Babiker, and M. Mayer. 2001. *Comparing greenhouse gases*. Cambridge, MA: MIT Joint Center for the Science and Policy of Global Change.

Richards, K., and R. Stavins. 2005. *The cost of U.S. Forest-based carbon sequestration*. Arlington: Pew Center on Global Climate Change.

Roughgarden, T., and S. Schneider. 1999. Climate change policy: Quantifying uncertainties for damage and optimal carbon taxes. *Energy Policy* 27: 415–429.

Schlenker, W., A. Fisher, and M. Hanemann. 2005. Will U.S. agriculture really benefit from global warming? Accounting for irrigation in the hedonic approach. *American Economic Review* 95: 395–406.

Schmalensee, R. 1993. Comparing greenhouse gases for policy purposes. *Energy Journal* 14: 245–255.

Schneider, S., and L. Goulder. 1997. Achieving low-cost emissions targets. *Nature* 389: 13–14.

Sedjo, R. 1995. The economics of managing carbon via forestry: An assessment of existing studies. *Environment and Resource Economics* 6(2): 139–165.

Smith, V. 2004. Fifty years of contingent valuation. In *The international yearbook of environmental and resource economics 2004–2005*, ed. T. Tietenberg and H. Folmer. Cheltenham: Edward Elgar.

Smith, J., and D. Tirpak. 1989. *The potential effects of global climate change on the United States: Report to congress*. Washington, DC: U.S. Environmental Protection Agency.

Tol, R. 2005. The marginal damage costs of carbon dioxide emissions: An assessment of the uncertainties. *Energy Policy* 33: 2064–2074.

Tulkens, H. 1998. Cooperation versus free riding in international environmental affairs: Two approaches. In *Game theory and the environment*, ed. N. Hanley and H. Folmer. Cheltenham: Edward Elgar.
Weitzman, M. 1974. Prices vs. quantities. *Review of Economic Studies* 41: 477–491.
Weitzman, M. 1998. Why the far-distant future should be discounted at the lowest possible rate. *Journal of Environmental Economics and Management* 36: 201–208.
Weyant, J., and J. Hill. 1999. The costs of the Kyoto Protocol: A multi-model evaluation, introduction and overview. *Energy Journal,* Special Issue.
Wigley, T., R. Richels, and J. Edmonds. 1996. Economic and environmental choices in the stabilization of atmospheric CO_2 concentrations. *Nature* 379: 240–243.
Williams, R. 2002. *Prices vs. quantities vs. tradable quantities*. Working Paper No. 9283. Cambridge, MA: NBER.
Yohe, G., H. Ameden, P. Marshall, and J. Neumann. 1996. The economic cost of greenhouse-induced sea-level rise for developed property in the United States. *Climatic Change* 32: 387–410.

Cliometrics

Robert Whaples

Abstract

Cliometrics (from Clio, the ancient Greek muse of history) studies history by applying the rigour of economic theory and quantitative analysis while simultaneously using the historical record to evaluate and stimulate economic theory and to improve comprehension of long-run economic processes. It thus allows mainstream economists to study economic history using their familiar methods. Since the 1950s, when cliometrics demonstrated that antebellum slave-owning was profitable, it has grown to become the dominant approach to economic history. It is now addressing traditional economic historians' topics like non-market behaviour and embracing methods and findings from disciplines beyond economics.

Keywords

Anthropometric history; Cliometrics; Computers in research; Domesday Book; Economic history; Fogel, R.; Neoclassical theory; North, D.; Slavery

JEL Classifications

N0

Cliometrics aspires to enhance the study of the economic past by applying the rigour of economic theory and quantitative analysis, while simultaneously using the historical record to evaluate and stimulate economic theory and to improve comprehension of long-run economic processes (Greif 1997). The term derives from Clio, the ancient Greek muse of history.

The methodology emerged in the United States in the late 1950s among a new generation of neoclassically trained economists who found that many historical writings contained analysis, frequently implicit, that did not conform to the minimum standards of economic literacy and so led to important misinterpretations of the historical record. Pioneering the use of computers in historical research, cliometricians were able to construct extensive macroeconomic time series and also to estimate economic relationships and marginal effects. Instead of imprecise qualitative statements such as 'it is difficult to exaggerate the importance of this', cliometrics tried to provide precise numerical estimates of economic magnitudes and economic relationships.

The potential value of the new approach was convincingly displayed in one of the first cliometric papers, Alfred Conrad and John Meyer's 'The economics of slavery in the Ante Bellum South' (1958). Earlier historians had wanted to compare the profitability of owning slaves with that of other investments, but didn't know how. Conrad and Meyer derived the average capital cost per slave, including the average value of the land, animals and equipment used by a slave. Estimates of gross annual earnings were generated from data on the price of cotton and the physical productivity of slaves. Net earnings were then obtained by subtracting maintenance and supervisory costs. The average length of the stream of net earnings was determined from mortality tables. The computation for female slaves took account of the

number and productivity of offspring, plus maternity and rearing costs. Conrad and Meyer's preliminary findings strongly refuted the dominant historical interpretation that slave owning wasn't profitable. Numerous subsequent refinements confirmed their conclusion, which is now almost universally accepted.

Among the early cliometric studies that transformed historical interpretation, several works stand out, including Douglass North's *The Economic Growth of the United States, 1790–1860* (1961), Robert Fogel's *Railroads and American Economic Growth* (1964), and Fogel and Stanley Engerman's *Time on the Cross*: *The Economics of American Negro Slavery* (1974). Indeed, Fogel and North's research was so influential that in 1993 the Royal Swedish Academy cited them 'for having renewed research in economic history by applying economic theory and quantitative methods in order to explain economic and institutional change', and awarded them the Nobel Memorial Prize in Economics as 'pioneers in … cliometrics' (Royal Swedish Academy of Sciences 1993).

One can gauge the rise of cliometrics by examining the *Journal of Economic History* (*JEH*). In the early 1950s fewer than two per cent of the pages in the *JEH* were devoted to cliometric articles, that is, those using extensive quantification and explicit economic theory. This figure subsequently climbed to ten per cent in the late 1950s, 16 per cent in the early 1960s, 43 per cent in the late 1960s, and 72 per cent in the early 1970s (Whaples 1991). In the late 1950s cliometrics was seen by some as a mere fad, but by the 1970s it was the standard approach for American economic historians. The cliometric tide has not ebbed; rather, the percentage of cliometric pages in the *JEH* rose to 83 per cent in the late 1980s and 90 per cent in 2004. Opening the pages of the *JEH, Explorations in Economic History* or the *European Review of Economic History* is very much like opening the pages of other empirically oriented economics journals, allowing mainstream economists to tackle historical research by familiarizing themselves with historical issues and applying the same methods they would use elsewhere. The overlap between cliometrics and economic history as practised by economists is now almost complete, as cliometrics has become dominant among economists doing historical research outside North America.

Cliometrics is not without critics. Traditional economic historians saw the young cliometricians as outsiders, as economists, not historians or economic historians; they claimed that these upstarts were theorists with little knowledge of the facts and with no sense of history, and that their findings were driven by restrictive theoretical assumptions (Goldin 1995). The economic historian had always been a hybrid, like the mule able to work in a challenging environment because it shared its parents' best traits. The cliometrician, on the other hand, wasn't a hybrid but was akin to a horse (or, worse, a jackass) that was trying to plough a field for which it was unsuited. Many historians found cliometric methods, models and multivariate regressions incomprehensible and could no longer keep up with research in economic history. Perhaps as a result many American history departments discontinued training and hiring specialists in economic history, and departments of economic history disappeared where they had been common outside the United States.

Many cliometricians, led by Douglass North, argued that most early cliometric research was too wedded to static neoclassical theory, which tends to focus analysis on historical episodes and topics for which markets were important but which severely limits the issues that can be examined. The neoclassical approach essentially assumes that the same preferences, technology and endowment lead to a unique economic outcome, implying that history does not affect equilibrium and that institutions other than the market don't matter. As the neoclassical grip was loosened in the 1980s, many cliometricians returned to studying issues traditional to economic historians, such as the nature and role of non-market institutions, culture, entrepreneurship, institutional innovation, politics, social factors, distributional conflicts, and the historical processes of economic growth and decline (Greif 1997). The field has also stretched its boundaries by taking seriously findings and methods from disciplines outside economics, such as the use of anthropometrics (which measures human stature and even skeletal remains) and by reaching

even further into the past, such as by analysing the efficiency of the English economy in the 11th century using data from the Domesday Book.

See Also

- ▶ Anthropometric History
- ▶ Economic History
- ▶ Fogel, Robert William (Born 1926)
- ▶ North, Douglass Cecil (Born 1920)

Bibliography

Conrad, A., and J. Meyer. 1958. The economics of slavery in the Ante Bellum South. *Journal of Political Economy* 66: 95–130.

Fogel, R. 1964. *Railroads and American economic growth: Essays in econometric history*. Baltimore: Johns Hopkins University Press.

Fogel, R., and S. Engerman. 1974. *Time on the cross: The economics of American Negro slavery*. Boston: Little, Brown and Co..

Goldin, C. 1995. Cliometrics and the Nobel. *Journal of Economic Perspectives* 9: 191–208.

Greif, A. 1997. Cliometrics after 40 years. *American Economic Review* 87: 400–403.

North, D. 1961. *The economic growth of the United States, 1790–1860*. New York: W. W. Norton.

Royal Swedish Academy of Sciences. 1993. The Sveriges Riksbank (Bank of Sweden) prize in economic sciences in memory of Alfred Nobel for 1993. Press release, 12 Oct Online. Available http://nobelprize.org/economics/laureates/1993/press.html. Accessed 27 June 2005.

Whaples, R. 1991. A quantitative history of the Journal of Economic History and the cliometric revolution. *Journal of Economic History* 51: 289–301.

Clubs

Suzanne Scotchmer

Abstract

The word ‘club’ has a deceptively frivolous connotation, as does the word ‘game’. But, like game theory, club theory has wide reach. By ‘club’ economists mean a small group of people sharing an activity, often in a context where they care about each other’s characteristics. Such activities may include production of goods and services (firms), production of education (schools, academic departments), sharing of private goods in small groups, and community life (churches, charity organizations). The formation of firms, choice of schools, and choice of games to play are all covered by club theory, as are social arrangements like marriage.

Keywords

Bundling; Capitalization; Club theory; Clubs; Competitive equilibrium; Congestion; Consumption; Education; Expected utility; Externalities; Games; Group formation; Group type; Land markets; Local public goods; Lotteries; Moral hazard; Multiple equilibria; Public goods; Schools as clubs

JEL Classifications

D71

Origins

The word ‘club’ entered the economics literature with a seminal (1965) paper of James Buchanan, who used it to describe a group of people sharing a public good. The key idea he introduced was that public goods are often subject to congestion, and in that sense exhibit some of the rival aspect of private goods. As a consequence, it may be more efficient to replicate a public facility for different (small) groups of users rather than to bear the congestion cost imposed by many people using the same facility. As we will see, club theory has subsequently developed to focus more on interactions among the members of a group, in particular, firms, than on the facilities they share, but both aspects are important.

Buchanan’s idea resonated with an idea of Tiebout (1956), who argued that ‘local public goods’ will be provided optimally if agents are free to choose among jurisdictions. He argued

that, if jurisdictions are relatively small, there should be enough jurisdictions and jurisdictional variety to satisfy most residents.

These papers led to the conjecture, pursued by a long list of scholars (see Scotchmer 2002), that competition should provide for optimal group formation. This was by analogy to other market contexts where demand and supply equilibrate at prices that support an efficient allocation, provided that all the actors, including firms, are small relative to the market. Allowing for group formation is a powerful extension of competitive theory, since groups have features that do not fit easily into the general equilibrium theory of Kenneth Arrow, Gerard Debreu and their successors. Such features include externalities among agents, learning of skills, and shared consumption of private goods, whether through rental markets or informal arrangements.

The research agenda surrounding clubs has only recently produced the modifications to general equilibrium theory that accommodate group formation. Along the way, it has been necessary to sort out competing equilibrium concepts, and the difference between models of pure group formation, for which I use the word 'clubs', and models of group formation where membership in the group is coupled to occupancy of land. For the latter I use the term 'local public goods.'

The distinction between clubs and local public goods is the focus of Scotchmer (2002), so I will not focus on it here. Local public goods economies differ from club economies in that jurisdictions are defined by geographical boundaries, and access to local public goods is intermediated by a land market. The price of land serves two related purposes: it allocates land within each jurisdiction, and in conjunction with capitalization effects, allocates agents among jurisdictions. An important complexity is that land and local public amenities are not generally priced or consumed separately. Instead, they are bundled. Although there are two price systems, local taxes and land prices, these cannot generally be interpreted as separate prices for local public goods and land, due to the bundling and to capitalization. In this environment, there are many possibilities for how to define a commodity space and price system, none entirely satisfying. The possibilities are more limited in the club model, where there is no land market that intermediates access to groups. Nevertheless, there are many nuances in adapting general equilibrium theory to group formation, which I now explore.

Clubs (Groups) in General Equilibrium

There have been two approaches to putting clubs into general equilibrium theory, which I refer to as the EGSSZ approach and the CPPT approach. The EGSSZ approach follows Ellickson et al. (1999, 2001, 2005, referred to here as EGSZ), Scotchmer (2005), Zame (2005), and Scotchmer and Shannon (2007). The CPPT approach follows Cole and Prescott (1997) and Prescott and Townsend (2006).

I begin with the EGSSZ model, and then discuss how it relates to the CPPT model. The commodity space begins with an exogenously given set of *group types*. In a state of the economy, there may be many copies of a given group type. A group type specifies a finite set of memberships, activities that the members engage in, and an input–output vector of private goods. Thus, group types may be interpreted as firms that produce private goods or use private goods as inputs to other activities. The memberships may have qualifications attached to them, such as to be smart or brawny, or to have skills such as the ability to write computer code. These qualifications are called *membership characteristics*. A given membership may or may not be available to an agent in his consumption set and, if it is, his qualification for the membership may be innate or learned.

Using the notation of Scotchmer and Shannon (2007), let $\mathbf{G}$ be a finite set of group types, and for each $g \in \mathbf{G}$, let $\mathbf{M}(g)$ be a set of memberships. Each membership $m \in \mathbf{M}(g)$ has attached to it a membership characteristic. The definition of the group type also specifies the group's activities and an input–output vector, say $h(g) \in \mathbf{R}^N$. Some group types do not require inputs or produce outputs; some require only inputs; and some (firms) may require inputs to produce outputs. Labour in a firm is not modelled as an input but rather as a

group membership for which skills or other characteristics may be required.

It is convenient to assume that a group's required input–output vector is distributed among members of the group. Thus, each group has associated to it an exogenously given transfer function $t_g : \mathbf{M}(g) \rightarrow \mathbf{R}^N$ such that $\Sigma_{m \in \mathbf{M}(g)} t_g(m) = h(g)$. The transfers specify each member's share of h, which may have positive and negative elements. Unless used for incentive purposes as in the papers referenced in section "Unverifiable Characteristics and Games" below, the transfer functions t_g can largely be arbitrary. Any maldistribution can be remedied through membership prices, discussed below, which are endogenous.

There is a continuum of agents, say, $A = [0,1]$. Each agent consumes a bundle of private goods $x \in R^N_+$ and a *list of memberships*, $\ell : \cup_{g \in G} \mathbf{M}(g) \rightarrow \{0, 1\}$. The value $\ell(m) = 1$ means that the agent chooses membership m, hence belongs to a group of type g such that $m \in \mathbf{M}(g)$. A state of the economy is (x_a, ℓ_a), $a \in A$, where $x_a \in \mathbf{R}^N_+$ is agent a's consumption of private goods and ℓ_a is a list of memberships. Each agent $a \in A$ has a utility function u_a, an endowment of private goods, $e_a \in \mathbf{R}^N_+$, and a consumption set. The utility function takes values $u_a(x_a, \ell_a)$, where $x_a \in \mathbf{R}^N_+$ is a consumption of private goods and ℓ_a is a list of memberships.

An agent's consumption set determines which memberships are available to him. For example, an agent's consumption set would presumably not permit both a membership in a sumo wrestling club and a membership in a ballet club, since the qualifications for those memberships cannot coexist in the same agent. Consumption sets play a much larger role in club theory than in private-goods economies. Some memberships may not be available to a given agent at all, regardless of what other memberships he chooses or what private goods he invests.

A state of the economy is feasible if it satisfies material balance in private goods, and if, in addition, membership choices are consistent with each other. Membership choices are *consistent* if there exist non-negative real numbers $\alpha(g)$, $g \in \mathbf{G}$, such that the number (measure) of agents who choose each membership $m \in \mathbf{M}(g)$ is $\alpha(g)$. Thus, $\alpha(g)$represents the number of type-g groups, and consistency implies that there are (almost) no groups that are only partially filled.

Consistency of membership choices presents the main technical difficulty in this model. The fixed point in the EGSZ (1999) proof of existence delivers prices such that membership choices are consistent. There is no analogous consistency condition for private-goods exchange economies, and consistency would typically be impossible if the club economy had a finite number of agents rather than a continuum.

To define equilibrium, we need two sets of prices: private-goods prices $p \in \mathbf{R}^N_+$ and membership prices $q : \cup_{g \in G} \mathbf{M}(g) \rightarrow \mathbf{R}$. The membership prices can be positive or negative. An agent's budget is determined by the value of his endowment and the value of the transfers he receives (or is obligated for) in his memberships, evaluated at the equilibrium private goods prices p. These must generate enough income to pay for his memberships at prices q and for his private goods consumption at prices p.

Stated informally, an equilibrium consists of private goods prices p, membership prices q, and an allocation (x_a, ℓ_a), $a \in A$ such that each agent is optimizing in his budget set; supply equals demand for private goods; the membership choices are consistent; and the membership prices sum to zero in each group type. Thus, the profit in each group is shared among the members – there is no notion of ownership of groups or group types.

Since the membership prices sum to zero within each group, some members pay other members. Intuitively, some members are paid because they create positive externalities or production opportunities for the members who pay. If, for example, there is a membership that relatively few agents are qualified to fill, or if it is costly to acquire the qualification, then that membership may have a negative price – the member is paid to belong to the group.

All the technical difficulties of general equilibrium theory appear here, such as the distinction between quasi-equilibrium and equilibrium. The technical difficulties in going from

C

quasi-equilibrium to equilibrium are exacerbated by group formation, since, for example, the inputs required for the group can exhaust the endowment of the members, who are then in the zero-wealth position. (See Gilles and Scotchmer 1997, example 3.)

I now give two informal examples of how club theory expands the reach of general equilibrium theory. First, let the group type be a firm that uses inputs to produce outputs. The required labour, with its required skills, is modelled through group memberships. The required skills might be innate for some workers, but for others might have to be acquired through investments of private goods or memberships in other group types, such as schools or apprenticeships. The negative elements of the input–output vector $h(g)$ are inputs, and the positive elements are the firm's output. These inputs and outputs are divided up among the workers (members) according to the transfer function t_g, and ultimately bought or sold in the market. The transfers contribute to the members' incomes. However, the income from the firm is further redistibuted through the endogenous prices (wages) q.

Substitution in the production process is modelled by using different firm types. If it is possible, for example, to produce the same input/output vector with many unskilled workers or with fewer skilled workers, those options would be modelled as different firm types. Whether a given firm type is used in equilibrium depends on the prices of private goods and memberships, the opportunity costs of workers (reflected in membership prices), and 'externalities' created within the firm type. Agents might avoid a very profitable technology because they dislike the production process or because they dislike the characteristics required of the other workers. This feature of production economies is not otherwise accommodated in general equilibrium theory.

Firms are perfectly competitive because each firm of a given firm type has measure zero in the economy, and therefore has no market power. Each firm makes zero profit even though there is no concept of linearity in production. The only constant returns to scale is that many copies of a given firm type may form, each producing the same output from the same inputs. However, each copy of the group type is a separate zero-profit entity.

Second, let the group type be a school. Suppose for simplicity that there are no private goods inputs or outputs, hence no internal transfers. Some of the memberships are called 'teacher' , and others are called 'student'. The same person is typically not qualified for both roles. The student memberships may be further differentiated. Some student memberships may be called 'advantaged student' (and require the appropriate qualification) and others 'disadvantaged student'. Which membership a student is qualified for is presumably constrained in his consumption set.

Since the membership fees sum to zero, the teacher will presumably be paid, and the students will pay. However, if advantaged students confer positive externalities on disadvantaged students, it might occur that both teachers and advantaged students are paid by disadvantaged students. Otherwise the advantaged students might prefer schools where all memberships are for advantaged students, where they themselves receive higher externalities.

The model I have described is a delicate amalgam of features inherited from the theory of general equilibrium for exchange economies and features of public goods economies, such as externalities and the sharing of private goods. In general equilibrium theory, the key features of a competitive equilibrium are that (*a*) the commodity space is defined independently of the set of agents, (*b*) the price system is complete with respect to the set of commodities, (*c*) prices are anonymous, and (*d*) agents optimize with respect to the price system, but not by observing other agents' preferences or endowments. Early discussions of price-taking equilibrium for club economies missed various of these requirements. For example, in analyses that use the 'core' equilibrium concept from game theory, following Pauly (1967), the commodity space has been defined as the set of groups (coalitions) that are feasible in the economy, even when the core is decentralized with prices. This idea departs from general equilibrium theory in that the available commodities

(group types) depend on the set of agents. That model has other limitations as well. Since agents can only belong to a single club, it cannot accommodate the notion that an agent may want to belong to several groups, for example, a school where he acquires skills and a firm where he exercises the skills. Further, many of the earlier models also restricted to a single private good (often with transferable utility), and therefore did not allow the important interpretation that groups are firms in a production economy.

In the model I have described, following EGSZ (2005) and Scotchmer and Shannon (2007), characteristics are defined as part of the membership, rather than attached to the agent. An agent can only choose a given membership if he is innately endowed with the characteristic required for it or, alternatively, can acquire it. The earlier models of EGSZ (1999, 2001) made the more restrictive assumption that all characteristics are innate, but the same proofs of existence of equilibrium and related theorems apply to both cases.

Randomized Memberships

In the model described above, agents choose memberships deterministically. However, the premise behind the CPPT branch of the clubs literature is that randomness can be utility enhancing, and randomness will therefore be created by the market. This depends on the premise that utility functions can be interpreted as von Neumann–Morgenstern utility functions (not assumed in the EGSZ model), and is illustrated by the following example.

Suppose there are two firm types, g_1 , $g_2 \in \mathbf{G}$. The firm type g_1 has a single worker and g_2 has a worker and supervisor. The club memberships are $\mathbf{M}(g_1) = \{m_{w1}\}$, $\mathbf{M}(g_2) = \{m_s, m_{w2}\}$. Suppose that each agent can choose a single membership, that a third of the agents have consumption sets that permit supervisor memberships, m_s, and two-thirds of the agents have consumption sets that permit worker memberships, m_{w1} or m_{w2}. There is a single private good, of which each agent has an endowment e. The utility of supervisors is equal to their consumption of the private good, regardless of memberships, and the utility of each worker is the following, where c is his consumption of the private good, and f is positive and increasing.

$$u(c,\ell) = \begin{cases} \frac{1}{2}f(c) & \text{if} \quad \ell = 0 \\ f(c) & \text{if} \quad \ell(m_{w1}) = 1 \\ f(c)+1 & \text{if} \quad \ell(m_{w2}) = 1 \end{cases}.$$

In an EGSZ-type equilibrium, the prices of memberships are $q(m_{w1}) = 0$ and $q(m_s) = -\hat{q}$, together with price $p = 1$ for the private good, where $f(e-\hat{q})+1 = f(e)$. Workers receive utility $f(e) = f(e-\hat{q})+1$ and supervisors receive utility $e+\hat{q}$. The supervisors are paid by the workers because agents who are qualified to be supervisors are relatively scarce and therefore valuable. They facilitate the creation of high value in supervised firms.

The basic idea of the clubs model of Cole and Prescott (1997) can be seen in the example. If the workers' utility function can be interpreted as a von Neumann–Morgenstern utility function, and if f is concave, the EGSZ-type equilibrium is not efficient. The expected utility of workers can be increased without decreasing the utility of supervisors by equalizing the workers' consumption in the two memberships m_{w1}, m_{w2}, and letting them randomize on those two memberships. The equalized consumption is $\hat{c} = (1/2)(2e-\hat{q})$. Then the *ex post* utility of workers who end up in m_{w1} is less than the *ex post* utility of workers who end up with m_{w2}, but their *ex ante* expected utility is the same, namely, $(1/2)f(\hat{c}) + (1/2)(f(\hat{c})+1) = f(\hat{c}) + (1/2)$, and larger than $f(e)$.

Cole and Prescott argue that the randomized outcome can be achieved in two ways. The agents can buy lotteries on club memberships directly, or the agents can buy randomizations on wealth and then choose their club memberships deterministically as in the EGSZ model. In the first implementation, prices are on units of probability placed on different consumption bundles. In the example, consumption bundles would be elements of some finite set $\mathbf{L} = \{(c,m)\}$, where, for mathematical convenience, c is in a finite set of points in $\mathbf{R}_+$ and $m \in \{m_s, m_{w2}, m_{w1}, m_o\}$, where m_o is a null membership that means no group membership is chosen.

The prices are $\{p(c, m) \in \mathbf{R}_+ : (c, m) \in \mathbf{L}\}$. If an agent chooses a consumption bundle (c,m) with probability one, he pays $p(c,m)$. More generally, an agent can choose probabilities (a 'lottery') $\{x(c, m) \in \mathbf{R}_+ : (c; m) \in \mathbf{L} \sum_{(c, m) \in \mathbf{L}} x(c; m) = 1\}$. It is then natural to define the utility function on the vectors x, so that the agent receives utility $u(x)$ and pays $p \cdot x$.

This transformation, also used by Prescott and Townsend (2006), gives the group-formation model a structure that is similar to an exchange economy. However, for analytical tractability some desirable features are given up along the way, such as that the authors assume there is a finite set of preference types, and restrict each agent to a single membership.

Moreover, there is a single profit-maximizing 'intermediary' on the supply side, which offers a combination of lotteries that maximizes profit, and creates firms from the outcome of agents' (independent) randomizations. To do this, the intermediary must serve a continuum of agents. The intermediary is therefore a different type of firm than the group types in the EGSSZ model and the firms of the CPPT model, such as g_1, g_2.

An important role of the intermediary in the CPPT model is to make transfers of value among the groups over which lotteries are offered. In the randomization above, the single membership in the firm type g_1 is coupled with consumption $\hat{c} < e$. The value of the member's consumption in g_1 is less than the value of the endowment, while the value of the members' consumptions in g_2 is more than the value of their endowments. Since the value of consumption must equal the value of endowments in aggregate, there is a transfer of value from g_1 to g_2. The intermediary who creates the lottery absorbs both sides of this transfer.

Scotchmer and Shannon (2007) show how lotteries on memberships can be introduced to the EGSSZ model through *lottery group types*, which are finite and are formally treated the same as ordinary group types. There is no need for a distinguished firm (intermediary) that serves a continuum of agents. A lottery group type is composed of several constituent group types in **G**. A feasible lottery must have the same number of lottery memberships as there are memberships in the constituent group types, since the lottery members will be assigned to the memberships in the constituent group types. The probability distribution is uniform on all assignments that are consistent with the memberships.

In the example, a lottery group type is constructed from one copy of g_1 and one copy of g_2, and has three memberships. Worker memberships to the lottery group type are such that the member can be assigned to m_{w1} or m_{w2}, and a supervisor membership is such that the member can only be assigned to m_s. There are two ways to make this assignment, each with probability one-half. Each worker has probability one-half of being assigned to mw_1 or mw_2, as required. If the lottery group type is defined such that the internal transfer of each worker to the supervisor is $e - \hat{c}$, the equilibrium membership prices q are zero.

With this structure, each lottery is a group type with finite memberships, and, as such, fits directly into the EGSZ model with no modification. Each worker pays the same membership fee for a lottery membership, but receives different *ex post* utility, depending on the outcome of the internal lottery. There are no transfers of value among lottery groups, as required by the zero-profit condition, but there are transfers of value among groups within each lottery group type.

A caveat is that not all lotteries can be accommodated with a finite number of group types. Each lottery group defines fixed probabilities on wealths and memberships. Different probabilities are provided by different lottery groups. Since there are continuously many possible lotteries, a complete lottery space would require a continuum of lottery group types, some very large. Thus, as in the CPPT approach, there is some loss in the technical convenience of restricting to a finite number of group types.

Unverifiable Characteristics and Games

In game theory the game is primitive. An agent either finds himself in the game or he does not, but there is generally no explanation for which game he finds himself in. Club theory allows agents to choose among games. However, to interpret a

game as a finite group type, the theory must accommodate strategies and characteristics that are not verifiable. Such an extension is suggested by Prescott and Townsend (2006), who use the CPPT approach to discuss how the market chooses among firm types that are subject to moral hazard. Equilibrium will weed out the contractual arrangements that are inefficient, where that may depend on the prices at which private goods trade. The same idea is taken up and extended by Zame (2005) and Scotchmer and Shannon (2007). The latter two papers build closely on EGSZ (1999, 2005) but differ in emphasis and in the way group formation is formalized.

Some unverifiable characteristics are chosen, and some are innate. The natural word for an unverifiable characteristic that is chosen is 'strategy', while it is more natural to say 'unverifiable characteristic' when the characteristic is innate but unobservable. Both play the same role in the model. In a normal-form game, for example, the membership might indicate row player or column player, and the strategy might indicate the unverifiable play. In a group type that is a firm, the membership is a job, and the unverifiable job characteristic might be innate proficiency at writing computer code.

When strategies (characteristics) are unverifiable, the groups that materialize from a member's choices will have a random component, namely, the unverifiable characteristics of other members. For random realizations of groups, Scotchmer and Shannon (2007) use the term 'augmented' group types. The agents first choose their verifiable memberships and unverifiable strategies, and are then randomly matched into augmented groups consistent with their choices.

If the unverifiable characteristics can be distinguished according to something verifiable like output, then group types can be defined such that agents screen optimally into groups, just as if the characteristics themselves were verifiable (see example 2 in Scotchmer and Shannon, 2007). No such ploy is available if the unverifiable characteristics affect utility directly.

After being matched into augmented groups, agents choose their consumptions of private goods. Each agent's income and demand for private goods may depend on the unverifiable characteristics in his groups. Since each agent's demand depends on the random matching, there is no conceptual reason to think that private-goods prices should be the same for all matchings, and Scotchmer and Shannon do not assume it. There may be two sources of uncertainty in an agent's consumption of private goods: uncertainty about the augmented groups and uncertainty about the prices of private goods. Both sources of uncertainty affect the *ex ante* demand for group memberships, and the optimizing choices of strategies.

If the set of agents were finite, the augmented groups realized by different agents could not be independent of each other. Duffie and Sun (2004a, b) show that the continuum remedies this problem. In the continuum, each agent's random match can be understood as independent of any other agent's random match, and a law of large numbers applies to demand. The law of large numbers provides an easy way to prove existence of equilibrium despite the randomness caused by unverifiable characteristics. If one assumes that the equilibrium prices must be the same at every random matching, aggregate demand can be treated as constant for all random matchings, and existence of equilibrium follows from EGSZ (1999). But this should not lead us to believe that constant prices are natural. There is no reason that the same equilibrium price vector should be selected at each random matching – constant prices are an assumption, not a conclusion. (This is an important difference between the treatments of Zame 2005, who assumes constant prices, and Scotchmer and Shannon 2007, who explore the consequences when prices can depend on the random matching. Variation in prices may reduce welfare.)

Prescott and Townsend (2006) prove the first welfare theorem for a class of economies with moral hazard. In contrast, Zame (2005) and Scotchmer and Shannon (2007) show many senses in which equilibrium will be inefficient. The difference lies partly in the classes of economies considered, and partly in the definition of 'efficiency', which is only defined relative to the trading opportunities in the economy. For

example, Scotchmer and Shannon point out that inefficiency in teams would vanish if agents could choose a game with a residual claimant. In the model of Prescott and Townsend, that is not an option.

These models have three broad classes of inefficiencies. First, the exogenous set of group types (games) in the economy may not be rich enough to achieve first-best efficiency, as in the teams example. Second, there are belief-driven coordination problems, well known in game theory, that are not solved by embedding games in general equilibrium. There may be multiple equilibria, including efficient ones and inefficient ones, each supported by beliefs that are correct in equilibrium. Third, there are inefficiencies in the trading of private goods. Trades in private goods are always efficient from an *ex post* point of view (conditional on the random matching) but not necessarily from an *ex ante* point of view. Depending on what is observable, the latter inefficiency may be remediable through insurance markets.

See Also

- ▶ Consumption Externalities
- ▶ Externalities
- ▶ General Equilibrium

Bibliography

Buchanan, J. 1965. An economic theory of clubs. *Economica* 33: 1–14.

Cole, H.L., and E.S. Prescott. 1997. Valuation equilibrium with clubs. *Journal of Economic Theory* 74: 19–39.

Duffie, D. and Y. Sun 2004a. *Existence of independent random matching,* Working paper. Graduate School of Business, Stanford University.

Duffie, D. and Y. Sun 2004b. *The exact law of large numbers for independent random matching*, Working paper. Graduate School of Business, Stanford University.

Ellickson, B., B. Grodal, S. Scotchmer, and W. Zame. 1999. Clubs and the market. *Econometrica* 67: 1185–1217.

Ellickson, B., B. Grodal, S. Scotchmer, and W. Zame. 2001. Clubs and the market: Large finite economies. *Journal of Economic Theory* 101: 40–77.

Ellickson, B., B. Grodal, S. Scotchmer, and W. Zame. 2005. The organization of consumption, production and learning. In *The Birgit Grodal symposium*, ed. K. Vind. Berlin: Springer-Verlag.

Gilles, R., and S. Scotchmer. 1997. On decentralization in replicated club economies with multiple private goods. *Journal of Economic Theory* 72: 363–387.

Holmstrom, B. 1984. Moral hazard and observability. *Bell Journal of Economics* 10: 74–91.

Pauly, M.V. 1967. Clubs, commonality and the core: An integration of game theory and the theory of public goods. *Economica* 34: 314–324.

Prescott, E.S., and R.M. Townsend. 2006. Firms as clubs. *Journal of Political Economy* 114: 644–671.

Scotchmer, S. 2002. Local public goods and clubs. In *Handbook of public economics*, ed. A. Auerbach and M. Feldstein, vol. 4. Amsterdam: North-Holland.

Scotchmer, S. 2005. Consumption externalities, rental markets and purchase clubs. *Economic Theory* 25: 235–253.

Scotchmer, S. and C. Shannon. 2007. *Verifiability and group formation in markets,* Working paper E07–347. Department of Economics, University of California, Berkeley.

Tiebout, C.M. 1956. A pure theory of local expenditures. *Journal of Political Economy* 64: 416–424.

Zame, W. 2005. *Incentives, contracts and markets – a general equilibrium theory of firms*, Working paper no. 843. Department of Economics, University of California Los Angeles.

Clusters

Charlie Karlsson
Jönköping University, Jonkoping, Sweden

Abstract

It is a fundamental structural characteristic of industrial economies that economic activities tend to co-locate, i.e. cluster in space. The study of clusters and clustering is today an integral part of many undergraduate and post-graduate studies in business administration, economics, economic geography, and urban and regional planning as well as a topic of research in these disciplines. At the same time, we can observe many governments at different levels in industrialized countries that have initiated cluster studies and introduced policies aiming at supporting existing clusters and stimulating the emergence of new clusters. The success of these policies has varied

substantially but cluster policies seem to have become an integral part of the industrial and regional policies in industrialized countries.

Keywords
Co-location; Agglomeration; Functional region; Cluster policy

JEL Classification
L26; L52; O12; O18; O25; R11; R12; R58

Introduction

It is a fundamental structural characteristic of industrial economies that economic activities tend to co-locate, i.e. cluster in space (Karlsson 2008a). The study of clusters and clustering is today an integral part of many undergraduate and postgraduate studies in business administration, economics, economic geography, and urban and regional planning as well as a topic of research in these disciplines. At the same time, we can observe many governments at different levels in industrialized countries that have initiated cluster studies and introduced policies aiming at supporting existing clusters and stimulating the emergence of new clusters. The success of these policies has varied substantially but cluster policies seem to have become an integral part of the industrial and regional policies in industrialized countries. There is currently a strong belief that clusters can be the major vehicle for economic development and growth in three ways: i) by increasing the productivity of the firms located in the cluster through internal and external economies of scale, ii) by increasing the pace of innovation through rapid knowledge exchange, and iii) by stimulating the formation of new firms, i.e. entrepreneurship, and the growth of firms (Huggins 2008).

The current large interest in clusters is a culmination of a research tradition that goes back to the late 19th century and is associated with names such as Marshall, Weber, Ohlin, Hotelling, Cristaller and Lösch. Even if these thinkers have contributed to the field over the years, it has been mainly economic geographers that have kept the research tradition running. Mainstream economists have largely ignored spatial issues until the early 1990s, when Krugman (1991) showed that the most striking feature of the geography of economic activity was concentration. It was first with the contributions by Krugman (1991) and Porter (1985) that research on clusters took off. While Porter mainly focuses clusters from a nation state perspective and how they generate competitive advantages, Krugman discusses clusters as something developing at the regional level due to specific centripetal forces that induce firms in individual industries to co-locate. According to Krugman, the geographic concentration of production is evidence for the pervasive influence of increasing returns. Since the early 1990s, a growing number of non-spatial economists have started to pay attention to what is known as "New Economic Geography". Fujita et al.(1999) note the increased theoretical and empirical interest among economists in where economic activities locate, why they concentrate in space and the importance of these processes for core areas in economics such as urban economics, location theory, and international trade theory.

Economic geographers mainly have accepted the economic analysis of clustering processes but stress that social, cultural and institutional factors also play an important role the development, growth and possible decline of clusters (Martin and Sunley 1997). The benefits of co-location are a function of the internal configuration of clusters. These benefits are derived from flexibility, informal networks based on frequent face-to-face interaction, trust-based interconnections among some large and many small firms and their subcontractors, specialized local infrastructures and institutions, a common skilled labour pool, and the rapid diffusion of knowledge and ideas (May et al. 2001). Perhaps the most important benefit of a cluster is information and knowledge advantage due to the potential for frequent face-to-face interaction that co-location generates. The information and knowledge is communicated through the 'industrial atmosphere' in the form of 'noise' and 'buzz'. The face-to-face interaction brings

distinct information including persistent updates, planned and unplanned learning, and the development of similar interpretation schemes, shared understanding of new knowledge and new technologies, which, over time develop shared cultural traditions and habits. However, the benefits do not come for free. Economic agents must establish and invest in links to other economic actors and build the necessary trust (Karlsson et al. 2005). Scott (1998) claims, for example, that clusters can only create new knowledge and new products and continue to grow if they also have linkages with external markets and utilise a mix of local and non-local transactions. Thus, the effects of local interaction and learning are much stronger if they are continuously supported by impulses from other regions and clusters (Bathelt 2005).

What is an industrial cluster and what do different researchers imply when using the concept? Despite substantial research on clusters, there is still much confusion concerning the proper conceptualization of a cluster, except that it is generally conceived as a non-random (Elison and Glaeser 1997) geographical agglomeration of firms, with shared complementary capabilities (Richardson 1972). Inside such clusters, one can identify several forms of intended and unintended interactions. Increasing returns occur when such interactions generate positive economic externalities for firms in the cluster. Gordon and McCann (2000) have offered some help by providing a comprehensive assessment of various theoretical frameworks where industrial clusters have been analysed. They stress that the phenomenon of industrial clustering has attracted researchers from several disciplines and research traditions employing a diverse set of conceptualisations, theoretical frameworks and analytical approaches, which has generated ambiguity. Concepts such as agglomeration, cluster, industrial district, regional economic milieu, and industrial complex have been used interchangeably often with very little concern about how to make them operational. Gordon & McCann identify three analytically distinct forms of spatial industrial clustering, each of them subject to a logic of its own:

- The classical model of pure agglomeration, referring to job matching opportunities and service economies of scale and scope, where externalities arise via the local market and local spill overs
- The industrial-complex model, referring to explicit links of sales and purchases between firms leading to reduced transaction costs
- The club model, also known as the social-network model, which focuses on social ties and trust facilitating cooperation and innovation

Whatever type of cluster, the phenomena of industrial clustering is evidence of the pervasive influence of increasing returns (Krugman 1991). Typical for clusters is the existence of one or several forms of direct and/or indirect interaction between economic agents. Increasing returns are obtained, when such interaction generates positive externalities for the economic agents in the cluster.

The three cluster notions above may coexist since local markets, local transaction links, and local social networks can be integrated in various combinations within functional regions. Thus, even if it is possible to analytically distinguish three "pure" cluster models, it is important to realise that industrial clusters often exhibit rich but complicated and integrated features, many of which may be difficult to create or influence by policy measures. Many industrial clusters are unique and are the result of specific historical circumstances. Cluster models give little guidance for the development of such clusters, since they are the result of specific circumstances, which are impossible to imitate.

The Home of Clusters: The Functional Urban Region

A functional urban region (FUR), characterised by its agglomeration of activities and by its intra-regional transport infrastructure, provides intra-regional proximity and facilitates an intra-regional mobility of people, products, and inputs. The basic characteristic of a FUR is the integrated

labour market, in which intra-regional commuting as well as intra-regional job and labour search is much more intensive than for the inter-regional counterparts (Johansson 1998). The border of a labour market region is a good approximation of the border of a FUR, which normally contains one employment centre but in some cases two or more employment centres.

Porter (2000, 254) defines a cluster as "a geographically proximate group of inter-connected companies and associated institutions in a particular field, linked by commonalities and complementarities. The geographic scope of a cluster can range from a single city or state to a country or even a group of neighbouring countries." One can question this kind of all-embracing definition. The concept of proximity totally loses its meaning due to Porter's fuzziness regarding the spatial boundaries of clusters. We prefer to name such concentrations beyond the functional economic region as industrial networks. The forces keeping such networks together are in several respects different from those keeping a cluster together within a single functional region.

The concept of market potential can be used to describe economic concentration and the opportunities of making contacts within and between such concentrations (Lakshmanan and Hansen 1965). There are strong reasons for making a precise distinction between the internal and the external market potential of a FUR. The geographic delineation of a FUR is in a fundamental way related to the identification of its internal market potential. The internal market potential is a measure of the market opportunities existing inside the borders of a FUR.

It is a common assumption in regional economics that products vary with respect to the contact or interaction intensity associated with their input and/or output transactions (von Thünen 1826). For products with standardised and routine transaction procedures, little or no direct contact between buyer and seller is necessary. Moreover, when the same supplier and customer repeat the same delivery, the interaction between these two actors can be routinized, and hence the contact intensity goes down, causing transaction costs to decline. However, many products are traded under complex (and contact-intensive) transaction conditions, which may involve many transaction phenomena such as inspection, negotiations and contract discussions, legal consultation and documentation of agreements. Such products may themselves be complex and have a rich set of attributes, but the basic thing is that from a transaction point of view, they are not standardised, and the interaction procedures are not routine. A special case of a contact-intensive transaction is when a product is customised and designed according to specifications by the customer in a process of supplier-customer interaction. Thus, we can assume that the contact-intensity associated with selling and delivering different products varies considerably.

Another common assumption is that interaction costs are much lower for transactions within a FUR than between FURs. This implies that contact-intensive products can be claimed to have distance-sensitive transaction costs and that these geographic transaction costs rise sharply when a transaction passes a regional border (Johansson and Karlsson 2001). This also implies that products can be distance-sensitive with respect to input transactions. Similar arguments apply to the labour market in the sense that individuals (firms) search for jobs (labour) mainly inside their FUR. Thus, the interaction frequency associated with distance-sensitive products supplied in each FUR including labour can be assumed to decrease with increasing (time) distance from the region's centre (Holmberg et al. 2003). It is a general result from spatial interaction theory, that the interaction intensity is a decreasing function of the time distance between origin and destination (Sen and Smith 1995).

For each type of product in a FUR, it is possible to divide the total market potential into the internal (intra-regional) and the external (inter-regional) market potential. Firms wanting to supply distance-sensitive products must find a sufficiently large demand for their sales inside their own FUR. When internal economies of scale prevail, the internal market potential must exceed a certain threshold if firms producing distance-sensitive products will be able to make

a positive profit, i.e. "economic density" matters (Ciccone and Hall 1996).

The size of the internal market potential in a FUR is among other things a function of its infrastructure provision. Interaction infrastructure offers high density combined with low transaction costs, i.e. a large accessibility (Johansson 1996). This implies that suppliers have a large accessibility to customers and that producers have a large accessibility to suppliers of specialised inputs and households supplying specialised labour inputs.

Infrastructure has two fundamental roles (Lakshmanan 1989): (i) it influences both the consumption and the production possibilities of societies, and (ii) it is intrinsically a collective good in the sense that it is common to all households and firms. Thus, infrastructure in a basic way influences the size of the internal and external market potential of a FUR by (i) extending its links for interaction trough space, and (ii) creating intra- and inter-regional accessibility. Infrastructure also extends over time through its durability, which creates sustainable conditions for production and consumption for extended time-periods.

The Emergence and Growth of Clusters

The traditional analyses of location and clustering using the resource-based theory of location and clustering (and trade) emphasise the relative abundance of resources "trapped" in a FUR (Ohlin 1933). Critical resources have the character of durable capacities, which on the one hand, include natural resources and on the other hand the supply of infrastructure in the form of facilities and networks, R&D organisations, existing production capacities with specific techniques, and the supply of different semi-immobile labour categories. Modern resource-based models often emphasise the supply of knowledge-intensive labour as a primary location factor. The durable capacities generate comparative advantages in the sense of Ricardo and influence the potential specialisation profile of a FUR. Although these characteristics are exogenously given in the short and medium term, a major part of the durable characteristics (except natural resources) change gradually over time and are created by investment and migration processes.

The resource-based approach has been challenged in recent decades by scale-based models (Dixit and Norman 1980; Krugman 1980; Ethier 1982), even though Ohlin (1933) explicitly refers to the role of scale economies nearly five decades earlier. These models explain location and clustering (and trade) in a context of internal and external economies of scale and local and external market potentials of FURs, where the dynamics of the interdependence between market size and economies of scale is essential. In the short and medium term, the properties of markets are durable phenomena, which create comparative advantages in the pertinent FURs. It is obvious that to understand the emergence and the growth and dynamics of clusters there is a need to bring the two approaches together. One possible approach to do this is to associate (i) the resource-based advantages to the input market potentials of each sector, and (ii) the scale-based advantages to the customer market potentials of each sector (Holmberg et al. 2003).

The realisation of scale economies and the associated potential of division of labour, i.e. decomposition of production, and specialisation are intrinsically related to the size of the accessible market (Stigler 1951). When the decomposition takes place within a firm, the firm takes advantage of internal economies of scale, and when decomposition leads to outsourcing of production, the firm may take advantage of external economies of scale. Internal economies of scale are technological phenomena related to individual firms and imply that productivity increases (the unit cost decreases) as output gets larger. They may be related to the existence of one or several productivity-enhancing indivisibilities (fixed-cost factors), such as indivisible equipment, knowledge resources including patents, brand names, material and non-material networks or set-up costs including learning how to do it (Koopmans 1957), i.e. a "catalyst", which must be present in the production process without being used up (Krugman 1990). It is not the absolute size of the fixed costs that matters. Instead, the

size of the fixed costs should be related to the size of the accessible demand (Chamberlin 1933).

In theories of agglomeration of firms, i.e. of clustering, internal economies of scale and the size of the internal and external market potential of FURs are used as the principal factors explaining the spatial agglomeration of firms. Internal economies of scale are essential components in all models, which emphasize the role of variety of outputs and inputs, respectively. Firms with internal economies of scale search for FURs with a large enough market potential for making it possible to produce with a profit and FURs in which many firms want to locate develop a large internal market potential. Some types of goods and many types of services are associated with large geographical transaction costs, implying that the intra-regional market potential determines whether profitable production is possible in a FUR or not. Thus, it is essential to classify products according to their distance sensitivity in terms of transaction costs. Based on such an approach one can identify specific categories of products with a potential to develop clusters in small, medium-sized and large FURs, respectively.

Industrial clustering cannot be explained solely by internal economies of scale. Of equal importance is the existence of external scale-economies, which are vital for a sustainable development of clusters in FURs. The first type of external economies of scale – localisation economies – is a systems phenomenon, which occurs when several firms, producing similar products, are in the same FUR. Localisation economies are vital for specialisation and clustering processes in small and medium-sized FURs (when they are not resource-based) (Johansson and Karlsson 2001). The second-type of external economies of scale – urbanisation economies – is another type of systems phenomenon, which occurs in large FURs hosting many different and interacting clusters.

The impact of external economies of scale in the form of location economies was emphasised already by Marshall (1920). A firm operating under constant returns to scale can benefit from positive external economies of the output from other firms in the same FUR, i.e. from external economies of scale (Chipman 1970). Localization economies generally play a central role in many models in urban and regional economics as well as in models of spatial product cycles (Mills 1967; Hirsch 1967).

According to Marshall's theoretical scheme, the positive industry-specific effects from clusters, i.e. the co-location of firms, have three sources, namely (1) non-traded local inputs, (2) local skilled-labour supply, and (3) information spillovers. The first category may be considered as distance-sensitive inputs. Due to high geographic transaction costs, these inputs are more expensive when delivered from sources outside the FUR. This implies that proximity becomes an advantage when firms are co-located, since the concentrated demand from the pertinent industry also attracts neighbouring firms, which are input suppliers. These input suppliers have their own internal economies of scale. Thus, it is important for them to have accessibility to a sufficiently large demand, which in this case is provided by the localised firms in the cluster. The desire of specialised input suppliers to be in the same FUR as their customers is determined by a combination of frequent interactions with their customers and distance-sensitive transaction costs.

The second category of localisation economies is related to a firm's labour acquisition costs. In a FUR, where a large share of the labour force already has specialised industry-relevant skills, the costs for a firm to expand its labour force may be lower than otherwise. For example, search and training costs can be assumed lower when the skilled labour pool is large in a FUR. At the same time, a cluster of firms can attract to the FUR a rich variety of labour categories, specialised to suit the industry in question. According to the above arguments, proximity to specialised input suppliers and specialised labour supply will imply that inputs can be acquired at lower total costs for given quality levels. Because of this, the described phenomena belong to the family of pecuniary externalities.

The third category, the information and knowledge available in clusters is a regionally available, semi-public good. This phenomenon has the character of a non-pecuniary externality, since it

brings benefits that are not charged at a price, except in the form of land prices. Information and knowledge are spread without being priced in the intra-regional neighbourhood, because in such an environment with intense face-to-face interaction it becomes prohibitively costly to privatise all information and knowledge. Hence, some of it will spill over, sometimes as the result of a conscious mutual exchange of information. The information and knowledge of importance concerns a wide area, such as information and knowledge about production technique, product attributes, input suppliers, customers, and/or market conditions. The Marshall approach provides an explanation of the sources of co-location economies within an individual industry in a single-industry cluster. Duranton and Puga (2004) describe Marshall's three mechanisms as sharing, matching and learning economies.

Another scheme for analysing agglomeration economies was outlined by Ohlin (1933). In contrast to Marshall, Ohlin focused more on how the individual firm is affected by co-location with other firms. In his classification, agglomeration economies have four origins:

- *Internal economies of scale* associated with the production technique of the individual firm
- *Localization economies*, which affect the individual firm as an influence from the industry to which it belongs
- *Urbanisation economies*, which arise from the size of the FUR and thus are external to the industry and its firms
- *Inter-industry linkages* of input-output type, where proximity to suppliers of intermediate inputs reduces their price.

Both input and customer market potentials tend to vary with the size of the FUR. This makes it possible to combine resource-based and scale-based models to explain the emergence and growth of clusters. We can assume that the larger the FUR, the larger the potential to combine internal and external economies of scale and the larger the economic density. Scale economies imply for large FURs a location advantage regarding all products with a "thin demand" and thus clusters in these industries mainly will be found in such regions. Thus, large FURs can specialise in "cluster diversity" and rely on the double forces of internal and external scale economies. However, scale economies constitute an equally important phenomenon for industrial clustering in FURs of all sizes. Also, smaller regions can develop a specialisation, i.e. a cluster, in a self-organised way, but in this case, the development is limited to a set of closely related products in the same industry with low geographical transaction costs supported by localisation economies.

The location of a firm to small and large FURs, respectively, may release a set of self-reinforcing circular processes, which in an endogenous change process give rise to one or several clusters through what Myrdal (1957) described as "cumulative causation." This form of positive feedbacks is in general constrained by the development of the demand in the FUR and in its external markets, and by the existing capacities in the form of built environment, accessibility based on transportation systems, production capacities, and labour supply. For certain activities, these constraints may not be binding, whereas other activities require adjustments of the durable capacities. The market potentials can be assumed to adjust at a faster time scale than the durable capacities. In a longer time perspective, the capacities and the economic milieu in a FUR will adjust through a system of coupled feed-back linkages. The interaction between scale economies and regional durable characteristics has the same nature both in small and large FURs, although external linkages to other (and larger) FURs are more vital in smaller FURs. For small and medium-sized FURs, the adjustment of durable capacities may be assumed rather specific given the narrow set of sectors, which form the specialisation nucleus of such regions. We may understand how the location of an individual firm may release a clustering process by referring to (i) a firm's customer market potential, (ii) a firm's input market potential, and (iii) a firm's labour-input market potential. In a similar manner, it is possible for the individual household to identify its (i) job market potential, (ii) housing market potential, and (iii) consumption market potential. The interaction

infrastructure will function as a support factor in the clustering process.

Clustering as an Entrepreneurial Process

Clustering processes are located and limited to the FUR where the initial entrepreneur or group of entrepreneurs decided to locate a new firm. The emergence of clusters is often triggered by events that make a natural or social asset of a FUR an important location factor for an industry or that encourage a local entrepreneur or group of entrepreneurs to engage in a specific industry (Feldman and Schreuder 1996). Entrepreneurs function as innovators when they transform new and existing knowledge into marketable products (Karlsson et al. 2014). They are also change agents and at the same time, as they are driven by the possibility to earn an entrepreneurial profit, they influence the conditions for other entrepreneurs to start and develop firms. They do this by changing the demand and supply conditions in the FUR over time and by developing norms and other informal institutions, which form the entrepreneurial climate in the FUR. Due to their co-location, firms are also able to develop trust-based relationships, not only with other firms in the same industry but also with other important economic agents in the FUR, such as suppliers, customers, public authorities, R&D institutions, and so on (Press 2006).

Cluster formation processes are not linear processes but can be described as adaptive, self-organising processes. These processes engage entrepreneurs as well as political decision makers and contribute to the establishment of supporting and governing functions as well as material and non-material infrastructures often with the help of public resources. This implies that the cluster and the regional specialisation created through the activities of entrepreneurs tend to become unique due to its history (Krugman 1991) and thus inherently difficult to copy (Feldman and Martin 2004).

When entrepreneurs during the cluster formation process decide to start new firms, they take advantage of those resources, which have accumulated over time, such as customer market potential, input supply potential, knowledge, and financial and social capital (Westlund 2006). Cluster growth is often driven by the start-up of "breakaway firms" (Jacobs 1969), i.e. firms started by entrepreneurs with experiences from the same industry. Entrepreneurs with experiences from the same industry create the cluster and contribute to its continued growth (Feldman and Romanelli 2006).

To the extent that these entrepreneurs are successful, their activities will further strengthen the economic milieu in the FUR including its knowledge base, institutions and social capital. Likewise, they increase the possibilities to take advantage of internal and external economies of scale and establish new firms. Successful clusters not only create their own resources, institutions, and potentials. They also attract resources, such as financial capital, labour and entrepreneurs from other FURs. However, there is no guarantee that clusters, which have developed well in early stages will continue to grow. There are examples of clusters, which after being successful in early stages start to deteriorate long before the mature stage (Feldman and Francis 2004).

Since entrepreneurs initiate economic activities and build up resources and market potentials, they are a necessary factor in the dynamic cluster formation process. Entrepreneurial processes are mostly localised processes. New firms are to a high extent started in the FUR where the entrepreneur works and has established commercial and social networks and has access to a customer market potential as well as an input supply potential.

Three Important Types of Clusters

The literature contains many empirical cluster studies ranging from case studies to general analyses of clusters within specific industries. Whatever industry we think of from food, textiles, metal manufacturing all the way to cars, ICT, restaurants and financial services we can observe clear tendencies to cluster. Due to the limited space, we here limit the discussion to three types of clusters: high-tech clusters, media clusters and financial clusters.

High-Tech Clusters

When explaining the clustering particularly of high-tech firms, it is natural to make a theoretical distinction between co-location forces working at the demand side (clusters offer a large enough demand for new distance-sensitive high-tech products) and at the supply side (clusters offer better conditions for creative and innovative activities).

Even if most studies of high-tech clustering have concentrated on supply side factors, it is worthwhile also paying interest to demand side factors. It seems clear that there is a strong tendency among high-tech clusters to be located primarily in those large FURs in the rich industrialised countries that often can be characterised as metropolitan regions. These regions offer good conditions for innovative firms developing new products, since they offer a large home market but also a high accessibility to the markets in other large FURs in the home country as well as abroad due to highly developed air traffic networks.

Large FURs are concentrations of company headquarters, company R&D divisions, other advanced industries, research universities, university hospitals, R&D institutes, and high-income earners, i.e. they are concentrations of demanding customers with a high willingness to pay for innovative products fulfilling their specific requirements. Thus, due to their demand structure, these regions are an excellent testing ground for new products. Due to their high internal accessibility, they offer good opportunities for extended periods of interaction with customers during the product development and testing phases. In other words, these regions offer a home market where new innovative products can be tested and nurtured in the first phase before exporting them to other large FURs and in the second phase more generally.

There are general incentives for entrepreneurs to locate their firms in large FURs because they are more likely to be better exposed to customers there. Searching is costly for customers who, *ceteris paribus*, will prefer to minimize search costs by purchasing in areas of concentrated supply. This is particularly relevant in markets with discerning potential customers with specific requirements, who are keen to enquire and search before placing a purchase order (Karlsson and Johansson 2006).

A further advantage of locating high-tech firms in large FURs is the positive information externalities in such regions, through which individual entrepreneurs and firms receive signals about the strength and content of regional demand by observing successful trades of established suppliers. Such observations also inform about varieties of existing products including lack of varieties, and trigger the development of new varieties. Moreover, the fact that a given firm is in a FUR with a successful high-tech cluster provides potential customers with an indication or image of quality.

Like all other entrepreneurs, high-tech entrepreneurs can reduce their risks by locating in large FURs (Mills and Hamilton 1984). To the extent that fluctuations in demand are imperfectly correlated across customers, the demand for products with high geographical transaction costs can be stabilized in such regions.

The concentration of purchasing power and demanding customers in large FURs is a stimulus to entrepreneurs to start imitating successful products and thereby often also improving them to take market shares from incumbents by being localised near them, i.e. in the same FUR (Hotelling 1929). Indeed, when the competition in the product market is imperfect, which is the case in high-technology markets generating many product varieties that are imperfect substitutes, geographical proximity increases competition in the product market (Fujita et al. 1999). The gain of such actions may be short-lived if further high-tech entrepreneurs enter, or if the incumbents in the region react to this unwanted competition. However, this kind of competition is critical to keep a high-tech cluster vital and vibrant, even if many high-tech firms over time may suffer from proximity to other firms and eventually fail.

On the supply side, large FURs offer high-tech entrepreneurs and firms advantages in terms of accessibility to a large pool of well-educated and specialised labour (Marshall 1920), particularly, specialised workers in different technical fields

but also in accounting, law, design, advertising, etc. This reduces the costs for starting, running and expanding new businesses (Krugman 1993). Furthermore, densely populated agglomerations are conducive to a greater provision of non-traded inputs, i.e. their producer service infrastructure is more developed. Such inputs are provided in greater variety, at lower costs and possibly at higher quality in large FURs (Krugman 1991). There also exist physical infrastructure benefits for high-tech entrepreneurs and firms to locate in such regions, in terms of access to highways and airports, and thus better accessibility to suppliers located in other regions at home or abroad.

However, and more importantly large FURs offer a concentration of and accessibility to R&D in companies, R&D institutes, and research universities, etc., as well as various arenas for knowledge diffusion and knowledge exchange. They also offer a high accessibility to knowledge generated in other large FURs by means of air travel, intra-company networks in large multinational firms, and networks of university researchers, which implies that they are well positioned to follow the knowledge developments in other FURs. Thus, large FURs offer advantages to high-tech entrepreneurs and firms in terms of a high potential to benefit from various and rich knowledge flows. This is particularly important when the knowledge is complex and tacit in nature (Jaffe et al. 1993). Generally, a form of informational externality accrues to new high-tech entrepreneurs from observing established firms that produces successfully in large FURs. For example, there are large potentials for product and production knowledge to spill over in large, dense FURs. Thus, the start-up frequency for any high-tech sector should increase with the existing density of firms in the actual sector. A final reason for advantages of large FURs for high-tech entrepreneurs arises from reductions in transaction costs (Quigley 1998). Search costs for customers, suppliers, services, and knowledge are lower in larger FURs. This implies that economies of information flows (Acs et al. 1992) on both the demand and the supply side are greater in large FURs than in smaller FURs. Thus, new high-tech firms are most likely to be started where the spillovers are greatest, and hence high-tech clusters are much more likely to emerge in large FURs than in small FURs.

In high-tech industries, a high share of the new ventures is started by former employees from incumbent firms using some of the technological know-how from their former employer (Klepper 2001). This implies, that existing high-tech firms characterised by a high level of technological know-how and continuous innovation provide a training ground for future high-tech entrepreneurs (Franco and Filson 2000). With mechanisms like this a high-tech cluster can secure renewal as well as continued growth for an extended period.

Media Clusters

One sector known for a strong tendency to cluster in large FURs is the media sector. There is today a growing literature dealing with media clusters and in particular with large media clusters, sometimes characterized as ‘global media cities’ (Krätke 2003). There are several obvious reasons to why media firms tend to cluster (Karlsson and Picard 2011). One reason is that many of products from the media sector, like films, are produced in the form of projects, which run for limited periods of time. Each such project needs to engage a large number of different specialists on a temporary basis. Only a relatively large cluster will offer a diverse enough supply of specialists to make such projects economically feasible. Another reason why media sector firms cluster is that many media industries are creative industries within or closely related to the cultural sector. The tendency of media and creative cultural industries to cluster has been documented often in the literature in recent decades (Karlsson and Rouchy 2015).

A parallel process to the emergence of large media clusters in large FURs has been the marked trend towards the globalization of several large media firms (Pratt 2000). The growth of media firms, not least through mergers and acquisitions has led to the formation of very large media groups, which not only occupy a prominent position in the cultural sector in individual countries, but also create increasingly global networks of branch offices and subsidiaries with presence in

many large media clusters (Krätke 2003). These large media groups tend to pursue a strategy that involves the integration and recombination of the media value chains at both the national and the global level. Another crucial strategy is the ambition to take advantage of diversification, i.e. the simultaneous or almost simultaneous distribution and exploitation of the same 'content' via different media, e.g. the print media, television programs and internet services. Furthermore, the globalization of media firms is related to the increased importance of intellectual property rights and copyrights. Copyrights provide the mean for controlling information and entertainment products and ensuring that they can be exclusively exploited in a national market (Bettig 1996).

In recent decades, the media sector has been strongly affected by technological changes. We can today observe that the developments within information and communication technologies (ICT) are stimulating the birth of new media services including the creation, manipulation and distribution of digital content (Gillespie et al. 2001). An interesting characteristic of these new services, which include software, databases, electronic libraries, new media, videos, broadcasting, etc., is that they do not just embody knowledge – they are knowledge and behave as such (Arrow 1962). These new services represent what Quah (1999) calls 'the weightless economy', i.e., an economy whose products are non-excludable, infinitely replicable and electronically transportable costless through space, like knowledge (Arrow 1962). This observation might lead to the conclusion that the location of the production of media products is a non-issue, since there are no raw materials that should be transported to the producer and no physical goods that should be distributed from producer to customers. Media firms could locate anywhere and the FURs would no longer host any clusters of media firms. So why do media firms continue to cluster in large FURs?

At the same time, as 'the weightless economy' develops, we can also observe tendencies of technology-related industrial convergence (Dosi 1988) in the emerging digital economy but also a breakup of old value chains (Ewans and Wurster 1997) followed by a new structuring of value chains, where takeovers and strategic alliances play a significant role (Hagel III and Singer 1999). There are today numerous claims that industries like telecommunications, computing and entertainment are converging and one day might evolve into one huge multimedia industry. This convergence might even have been increasing in recent years with the emergence of the Internet and with the increasing capability of existing networks to carry both telecommunications and broadcasting services (Knieps 2003). Developments in digital technologies and software are creating a large technological innovation potential for the production, distribution and consumption of information services. Convergence, characterised as the ability of different network platforms to carry essentially similar kinds of services, may have very different faces: telecommunications operators may offer audio-visual programming over their networks; broadcasters may provide data services over their networks, cable operators may provide a range of telecommunication services, etc.

In the 1980s and early 1990s, some cyber prophets and technological optimists predicted that the emergence of the digital economy would kill distance and make clustering in large FURs superfluous (Cairncross 1997) and at the same time eliminate the scale disadvantages of smaller and more peripheral FURs. The basic idea was that the spread of the use of ICT has the potential to replace face-to-face activities that formerly occurred in central FUR locations, which would strongly reduce or even eliminate agglomeration economies and hence make economic activities totally 'foot-loose'. At the beginning of the 21st century, however, it has become clear that this picture is at least single-sided. New technologies are likely to remain grounded in existing FURs, implying that these regions will keep their locational attractiveness and that media clusters will remain or even grow. Thus, the ICT has not rendered work and organisation 'space less' (Neff 2005).

There is also increasing evidence that the digital revolution reinforces the position of leading FURs (Castells 1996). So why do media firms

cluster, when the technological opportunities have seemingly reduced the necessity of proximity in operations between inter-linked firms? It seems as if the clustering tendencies are even more dominant in media industries than in many traditional industries. Ogawa (2000) shows, for example, that ICT development may not necessarily encourage the dispersal of economic activities due to the network effects and the technology effects of ICT infrastructure supply. FURs are a means of reducing the fixed travel costs involved in face-to-face interactions. Even if in principle improvements in ICT could eliminate the demand for face-to-face interactions and make large FURs obsolete, empirical results point in the direction that telecommunications are mainly a complement to face-to-face interactions (Gaspar and Glaeser 1998).

A major effect of the rapid diffusion of ICT is a dramatic reduction in transport and communication costs, which will alter the incentives for clustering of media industries and firms. It is too early to observe the results of the diffusion of ICT but it is possible to identify some possible effects (Venables 2001), since ICT reduces

- the search and matching costs in product markets but closeness by customers may still be essential, for products with rich and fluent characteristics
- the direct shipping costs since many products can be delivered in digital form,
- the control and management costs for geographically and organisationally fragmented operations
- the cost of time in transit, i.e. the shipping to and communication with distant locations
- the costs of personal interactions and stimulates knowledge spillovers
- the costs of commuting and of travelling in agglomerations
- the costs of replicating products, and
- the costs of relocation

These effects are not specific for the media sector. However, due to the character of the media sector's products, ICT might have stronger effects on the media sector than on other sectors. It is by no means clear how these factors will affect clustering in the media sector even if it is obvious that the ICT revolution makes it possible for media firms to go from a physical to a virtual value chain as well as to eliminate stages in the value chain (Ghosh 1998). This implies that it is an empirical issue to find out how the clustering in the media sector as well as its different industries is affected by the ICT revolution.

Financial Clusters

The financial services industry has as its primary function to intermediate between savers and borrowers in an economy. Its secondary function includes financial management, risk pooling and facilitation of payments and the transfer of money. Concerning the location of financial services, it is fundamental to make a distinction between retail and wholesale financial services. Retail financial services offer services such as payment systems, savings accounts and loans to the general public and to small and medium-sized firms. Even if these services increasingly are handled over the Internet, the location of retail financial services offices tend to follow the spatial distribution of population and small and medium-sized firms, i.e. the customers of these services.

Wholesale financial services include issuing of bonds and equities, support to mergers and acquisitions, and sophisticated products for managing risks, such as financial derivatives and these services are offered to large companies not least multinational companies and governments. Wholesale financial services have always been concentrated in major cities. Europe's first bankers in the late middle ages located their operations in the major trading centres. The leading commercial centres also developed into centres for banking, insurance and other financial services. During the industrial age growing industrial production and international trade stimulated the concentration of wholesale financial activities to cities with a strategic location in the global networks, i.e. gateways, such as London, New York and Tokyo (Kindleberger 1974; Andersson 2000; McCann and Acs 2011). In recent decades, the wholesale financial services industry has consolidated by means of mergers and acquisitions

(Amel et al. 2004) at the same time as the wholesale financial markets have become global after the deregulation of financial markets in many countries (Slager 2006).

In the developed part of the world, we can today observe that wholesale financial services are concentrated in financial clusters in a small number of metropolitan functional regions (FURS) (Bindemann 1999; Poon 2003). A financial cluster can be described as "the grouping together, in a given urban space, of a certain number of financial services" or as "the place where financial intermediates coordinate transactions and arrange the settlement of payments" (Cassis 2006, 5). This clustering of wholesale financial firms has been driven by a most radical transformation of the global economy in terms of a rapid growth and international transactions in currencies, shares, bonds and other types of securities. This transformation has been made possible through the development of efficient international communication networks that have been expanding to ever-increasing capacities through digitalization and the Internet. Households, pension funds and other financial investors have experienced an increasing international accessibility to different financial markets coupled with increasing economic efficiency and generally lower transaction costs. By increasing the international diversification of portfolios, it has been possible for investors to increase the returns on their investments while keeping the risks constant.

Begg (1992) has developed a typology of such wholesale financial services clusters that can be a useful starting point for a discussion:

1. The first-order clusters consist of three global financial clusters that play in a class of their own: London, New York and Tokyo. These diversified clusters are where the headquarters of major financial services firms and institutions are located and they offer a wide range of financial services (Andersson 2000).
2. The second-order clusters are also diversified but they serve not the whole global economy but supra-regional parts of the global economy. These clusters include Hong Kong, Frankfurt, Paris, Sidney and Singapore.
3. The third-order clusters are specialist or niche international clusters such as Amsterdam, Edinburgh, Luxembourg, Boston, San Francisco, Washington DC, Toronto, Geneva and Zurich.
4. The fourth-order clusters are mainly national centres with more limited involvement in international business and include cities like Stockholm, Rome, Milan, Hamburg, Dublin and Barcelona.

Which are then the general centripetal forces that can be supposed to stimulate the emergence, growth and concentration of large financial services clusters? We can identify a number of such forces (Abraham et al. 1994; Pandit et al. 2008):

- Capitalization and high liquidity
- Macroeconomic conditions
- Opportunities to co-locate near competitors and a variety and critical mass of complementary firms and related services, including financial consultants and other advisers, business journalists, rating offices, analysts, traders, corporate financial officers, bankers and financial regulators,
- Access large pool of skilled and flexible labour,
- A large market size in terms of size and scope,
- Access to financial infrastructure and a highly developed financial market from a technological point of view,
- Access to physical infrastructure, in particular, office space, telecommunications networks and international transport links with enough capacity, not least for the interaction with customers that are external to the cluster,
- Access to localized information, knowledge and technology spillovers with a potential to generate financial innovations.
- A political organization and tradition with 'thick' supportive institutional structures, including regulatory solutions and tax system,
- Low transaction costs,
- The 'right' initial conditions including a financial tradition.

Financial clusters tend to have a very central location in the core of a metropolitan FUR. This

agglomeration of financial activities is explained by the critical role of external economies of scale in financial markets. Not least are there substantial economies of scale in the gathering and analysis of information about scientific and applied R&D and product development in R&D-intensive and high-tech industries and firms. Of course, it is impossible for most financial investors to gather and professionally analyse such information. Instead there has been a steady growth of consultancy firms and units within investment banks and other financial actors specializing in the analysis of such information. The specialists working with information gathering and analysis are dependent upon frequent participation in scientific conferences, informal discussions with scientists, patent engineers and other experts on scientific and high-tech innovation and diffusion. Economies of scale in the analysis of information and knowledge related to R&D and innovation are becoming an increasingly important factor for the efficiency of financial clusters. The critical role of economies of scale in wholesale financial activities is clearly illustrated by the fact that even the largest countries only host a handful of financial centres. The efficiency of a financial cluster depends on its internal connectivity, its capacity to communicate internally and globally and the knowledge base of the regional financial network.

In terms of location, i seems as if the specific address within this central location is important for the wholesale financial services firms. To be considered a serious player in the financial market it is important to have the right address, since such an address signals a strong brand. Personal contacts and opportunities to interact frequently face-to-face within walking distance are of critical importance for the functioning of a financial cluster, since face-to-face meetings involving many persons have the advantage that more complex information can be transferred as well as non-verbal signals indicating the degree to which different partners are happy with the agreement discussed. Also, the opportunity to have face-to-face meetings with regulators is important for wholesale financial services firms in a financial cluster. The customers demanding advanced financial services on their side appreciate the opportunity to find many suppliers of similar and related financial services in large financial clusters and value the opportunities to compare the products and prices of many suppliers of financial services concentrated in a limited geographical area. Customers of advanced financial services look for high quality suppliers and react to market signals such as the reputation of different financial clusters and different wholesale financial services firms.

Labour with financial skills are attracted to large financial clusters because the size of the market offers higher chances of continuous employment. A larger market also gives incentives for people to develop highly specialised financial skills. Furthermore, the movement of people between the financial firms in the cluster helps to develop a network of contacts and facilitates the diffusion of information and knowledge spillovers.

Cluster Policies

Cluster policies are currently a hot topic. Policy makers in many countries at both the national and the regional level have come to believe that supporting and creating clusters is the major industrial policy option to be competitive and to be a winner in the globalisation race (Lundvall 2002). Certainly, there is a strong need for a thorough discussion of cluster policies and not least the rationale for cluster policies. At different levels in many countries, cluster development has become **the** solution to economic development. However, in many cases, cluster development policies seem to be based on no or very limited analysis. Clusters are found and identified without any clear objective criteria. When criteria are used, they are often very simple, such as location quotients. Still worse, there is often very little analysis of what factors that gave rise to the emergence of different clusters, the factors keeping them together, the long-term prospects of the clusters, the fundamental reasons motivating political intervention, and the problems of applying cluster policies (Karlsson 2008b).

Existing clusters can possibly but without certainty be supported by policies. Stimulating the emergence of new clusters is substantially more complicated. Having witnessed the success of a limited number of successful high-tech clusters, many regions want to initiate and nurture their own high-tech clusters. This is often done with little and mostly superficial analysis. Often the initiatives to create new clusters are based upon rather simple imitation strategies, which severely underestimate the difficulties of launching new clusters. The difficulties are real since research has rather little help to offer to identify the necessary and sufficient conditions for successful launching of new clusters.

Clusters contribute positively to real income levels in FURs. This has important implications for regional development polices. However, it is not obvious what the implications are and how cluster policies should be designed (Karlsson and Stough 2002). What type of regional cluster policy to apply depends on

1. the types of cluster(s)
2. the actual degree of cluster formation in the FUR, and
3. the information and knowledge about existing clusters and potential cluster policies possessed by relevant political authorities.

In the ideal case, policy measures should be directed towards the causes of the problem to be solved. It is important to realise that externalities, which stimulate cluster formation is a sign of what is called a market failure. This holds irrespective of whether the externalities are pecuniary or non-pecuniary. In traditional economic welfare theory, the existence of market failures has generally been taken as a motivation for political interference. However, this view has become more nuanced in recent decades. Political interference is associated with its own costs and these costs must be weighed against the benefits from removed or reduced market failures.

In the case of non-pecuniary externalities, market failure is obvious. The individual firm has no incentives in its calculations to consider the positive (negative) effects for other firms in the cluster of its own activities. This condition implies, for example, that private firms in a cluster regularly under-invest in R&D, because they do not consider the value for other firms of its knowledge creation.

Pecuniary externalities on the other hand are market failures connected with scale economies or imperfect competition. The utilization of scale economies, the supply of products, and the degree of competition are all limited by the size of the accessible market potential. If more customers enter the market or if suppliers can have better access to distant markets, the scale limitation is reduced and socio-economic benefits accrue through lower unit costs, coupled with a wider supply of products and/or increased competition. Thus, it is not the pecuniary externalities as such, which represent market failure. It is just a symptom of a market failure, which comes from the production conditions (scale economies) or the market form (imperfect competition).

Certain market failures due to externalities can be avoided if the effects are internalised, e.g. if the firms in a cluster decide to coordinate their activities through a common ownership or through contractual arrangements. Cluster firms can also organise themselves and work jointly to get more firms and/or households to locate in the FUR to increase its market potential, if the size of the market potential is too small for all potential positive pecuniary externalities to be realised. There are, in fact, plenty of examples of the role that private sector leadership can play for cluster initiation and cluster development (Stimson et al. 2002). However, if the number of economic actors is large it might be impossible to achieve internalisation or to organise a private sector leadership. There are also limitations to what cluster firms can achieve on their own. Many important policy issues, such as the building up of material and non-material infrastructures, in most countries reside within the public sector. Obviously, there are two cases when public sector cluster policies might be considered under assumptions of perfect information. The first case concerns private sector coordination failures, where private sector coordination might be substituted with public sector coordination. The second case concerns

sub-optimal market potentials in FURs with clusters, where public sector infrastructure investments can contribute to increased market potentials by means of the geographical extension of FURs and/or better access to external markets. Coordination failures and/or under-optimal market potentials can result in clusters operating under suboptimal scale or that potential profitable clusters are not established.

As a cluster consist of those firms, which best can take advantage of the market potential in a FUR and its durable resources, cluster policies at the FUR level if anything should primarily focus on supporting and developing existing clusters. Due to the existence of positive externalities, the existing clusters in a FUR will normally not achieve an optimal scale spontaneously. However, to the extent that existing clusters are not capable of driving the economic development in a FUR, it might be natural to raise questions about new clusters and thus the possibilities for structural change in a FUR through cluster substitution (Venables 2001).

Even if there might exist basic welfare arguments for cluster policies, there is still the underlying problem that the relevant authorities in a FUR often lack the necessary information and knowledge about

- the character of the cluster benefits
- what the exact causes of the cluster benefits are
- which clusters that generate particularly strong cluster benefits
- what constitutes the coordination problem, and
- the role of intra- and interregional market potentials for clusters

Furthermore, there are other problems related to cluster policies, such as the risks for manipulation and lobbying, and the existence of asymmetric information.

Another problem related to cluster policies is that different economic processes work at different time scales. Product markets, for example, normally change through relatively rapid processes, which generate demands that durable regional characteristics, such as the labour force with its pertinent skills, real capital, infrastructure capital, and so on, should be adjusted. As dynamic competition drives many relatively rapid processes, there is a constant need to upgrade the economic milieu in FURs with clusters. The problem is that such capacity and quality adjustments are a much slower and above all a more sluggish process than the processes in the product markets (Johansson and Karlsson 2001). If the lags in the development of labour supply, environment and infrastructure are large, the growth of clusters may be retarded and rapidly turn into a negative phase. The possibilities to counteract lags in the capacity and quality adjustments and to create conditions for a sustainable cluster growth rests in long-term and credible cluster policies in a FUR that can reduce uncertainty among economic actors about the future growth prospects of the different clusters.

According to the modern theory of endogenous, regional growth, cluster growth is something that grow out of internal regional conditions that are susceptible to change (Johansson et al. 2001). In line with this view, cluster policies deal with conditions, which essentially must be developed and implemented with region-specific knowledge as a base. Thus, cluster policies if they shall be implemented must be implemented at the FUR level even if a more comprehensive view and financial support might come from the national level.

Internal economies of scale mainly rest outside the domain of economic policies. However, policies, which lead to lower fixed costs for labour and capital, reduce the dependence of firms on the size of the market potential in a FUR. Moreover, to get new and growing clusters running, it is important to create optimal conditions for start-ups, spin-offs, spin-outs, and firm growth. It is also important to create a clear vision and strong image for new clusters by means of a conscious and profiled marketing.

The geographical transaction costs are partly determined by the infrastructure and transport policies, which in many countries are determined at the national level. Lower geographical transaction costs extends the borders of FURs and increases their market potential, which creates scope for the development and growth of more industries and clusters and of firms with internal economies of scale. Transport costs are becoming an increasingly

important factor for the development of clusters as other costs connected to international trade decrease. It is important to observe that the profitability of investments in infrastructure is larger in FURs with clusters than in FURs without clusters. Normally cost-benefit calculations of infrastructure investments use to disregard this.

Large parts of the knowledge generation in a FUR is characterised by collective characteristics. Knowledge developed by one firm tends over time to diffuse to other firms in the FUR. This generates increasing returns in the FUR economy i.e. the growth of the FUR economy can be influenced by investments in knowledge, R&D, and human capital. Even if there is no one-to-one-relationship between knowledge-intensity and profitable clusters, there are still strong reasons to believe that clusters are more common in knowledge-intensive industries than in other industries. This implies that if a FUR wants to stimulate cluster growth and cluster formation, there are strong reasons for public investments in higher education and R&D. However, it is important to notice that precision in this case is more important than volume. The investments in higher education and R&D must be cluster relevant.

See Also

- ▶ Location Theory
- ▶ Spatial Economics
- ▶ Urban Production Externalities
- ▶ Urban Agglomeration

Bibliography

Abraham, J.-P., N. Bervaes, and A. Guinotte. 1994. The competitiveness of European international financial centres. In *The changing face of European banks and securities markets*, ed. J. Revell, 229–277. Basingstoke/London: Macmillan Press.

Acs, Z.J., D.B. Audretsch, and M.P. Feldman. 1992. Real effects of academic research: Comment. *American Economic Review 82*: 363–367.

Amel, D., C. Barnes, F. Panetta, and C. Calleo. 2004. Consolidation and efficiency in the financial sector. A review of international evidence. *Journal of Banking and Finance 28*: 2493–2519.

Andersson, Å.E. 2000. *Gateways to the global economy*. Cheltenham: Edward Elgar.

Arrow, K.J. 1962. Economic welfare and the allocation of resources for invention. In *The rate and direction of inventive activity*, ed. R.R. Nelson, 609–625. Princeton: Princeton University Press.

Bathelt, H. 2005. Cluster relations in the media industry: Exploring the 'distanced neighbour' paradox in leipzig. *Regional Studies 39*: 105–127.

Begg, I. 1992. The spatial impact of the completion of the EC internal market for financial services. *Journal of Regional Studies 26*: 333–347.

Bettig, R.V. 1996. *Copyright culture*. Bolder: Westview.

Bindemann, K. 1999. *The future of european financial centres*. London: Routledge.

Cairncross, F. 1997. *The death of distance*. Boston: Harvard Business School Press.

Cassis, Y. 2006. *Capitals of capital*. Cambridge: Cambridge University Press.

Castells, M. 1996. *The rise of the network society. The information age: Economy, society and culture*. Vol. 1. Oxford: Blackwell.

Chamberlin, E. 1933. *The theory of monopolistic competition*. Cambridge, MA: Harvard University Press.

Chipman, J.S. 1970. External economies of scale and competitive equilibrium. *Quarterly Journal of Economics 72*: 347–385.

Christaller, W. 1933. *Die zentralen Orte in Süddeutschland*. Jena: Gustav Fischer.

Ciccone, A., and R.E. Hall. 1996. Productivity and the density of economic activity. *American Economic Review 86*: 54–70.

Dixit, R., and V. Norman. 1980. *Theory of international trade*. Cambridge: Cambridge University Press.

Dosi, G. 1988. Sources, procedures and microeconomic effects of innovation. *Journal of Economic Literature 36*: 1126–1171.

Duranton, G., and D. Puga. 2004. Micro-foundations of urban agglomeration economies. In *Handbook of regional and urban economics, volume 4, cities and geography*, ed. J.V. Henderson and J.-F. Thisse, 2063–2117. Amsterdam: Elsevier.

Ellison, G., and E.L. Glaeser. 1997. Geographic concentration in U.S. manufacturing industries: A dartboard approach. *Journal of Political Economy 105*: 889–927.

Ethier, W. 1982. National and international returns to scale in the modern theory of international trade. *American Economic Review 72*: 389–405.

Ewans, P.B., and T.S. Wurster. 1997. Strategy and the new economics of information. *Harvard Business Review 78*: 71–103.

Feldman, M.P., and J. Francis. 2004. Homegrown solutions: Fostering cluster formation. *Economic Development Quarterly 18*: 127–137.

Feldman, M.P., and R. Martin. 2004. *Jurisdictional advantage*, *NBER working paper* No. 10802.

Feldman, M.P., and E. Romanelli. 2006. Organization legacy and the internal dynamics of clusters: The U.S.

human bio-therapeutics industry, 1976-2002, Paper presented at the 2006 DRUID Winter Conference
Feldman, M.P., and Y. Schreuder. 1996. Initial advantage: The origins of the geographical concentration of the pharmaceutical industry in the Mid-Atlantic region. *Industrial and Corporate Change 5*: 839–862.
Franco, A.M., and D. Filson. 2000. *Knowledge diffusion through employee mobility*, *working paper.* Iowa City: University of Iowa.
Fujita, M., P. Krugman, and A.J. Venables. 1999. *The spatial economy: Cities, regions and international trade*. Cambridge, MA: The MIT Press.
Gaspar, J., and E.L. Glaeser. 1998. Information technology and the future of cities. *Journal of Urban Economics 43*: 136–156.
Ghosh, S. 1998. Making business sense on the internet. *Harvard Business Review 76*: 126–133.
Gillespie, A., R. Richardson, and J. Cornford. 2001. Regional development and the new economy. *European Investment Bank Papers 6*: 109–131.
Gordon, I.R., and P. McCann. 2000. Industrial clusters: Complexes, agglomerations and/or social networks. *Urban Studies 37*: 513–533.
Hagel, J. III, and M. Singer. 1999. Unbundling the corporation. *Harvard Business Review 77*: 133–141.
Hirsch, S. 1967. *Location of industry and international competitiveness*. Oxford: Oxford University Press.
Holmberg, I., B. Johansson, and U. Strömquist. 2003. A simultaneous model of long-term regional job and population changes. In *The economics of disappearing distance*, ed. Å.E. Andersson, B. Johansson, and W.P. Andersson, 161–189. Aldershot: Ashgate.
Hotelling, H. 1929. Stability in competition. *Economic Journal 39*: 41–57.
Huggins, R. 2008. The evolution of knowledge clusters: Progress and policy. *Economic Development Quarterly 22*: 277–289.
Jacobs, J. 1969. *The economy of cities*. New York: Vintage.
Jaffe, A., M. Trajtenberg, and R. Henderson. 1993. Geographical localisation of knowledge spillovers as evidenced from patent citations. *Quarterly Journal of Economics 108*: 577–598.
Johansson, B. 1996. Location attributes and dynamics of job location. *Journal of Infrastructure Planning and Management 530*: 1–15.
Johansson, B. 1998. *Infrastructure, market potential and endogenous economic growth*, paper presented at the Kyoto workshop 1997, Department of Civil Engineering, Kyoto University.
Johansson, B., and C. Karlsson. 2001. Geographic transaction costs and specialisation opportunities of small and medium-sized regions: Scale economies and market extension. In *Theories of endogenous regional growth – lessons for regional policies*, ed. B. Johansson, C. Karlsson, and R.R. Stough, 150–180. Berlin: Springer.
Johansson, B., C. Karlsson, and R.R. Stough, ed. 2001. *Theories of endogenous regional growth – lessons for regional policies*. Berlin: Springer.
Kaldor, N. 1970. The case for regional policies. *Scottish Journal of Political Economy 17*: 337–348.
Karlsson, C., ed. 2008a. *Handbook of research on cluster theory.* Cheltenham: Edward Elgar.
Karlsson, C., ed. 2008b. *Handbook of research on innovation and clusters. Cases and policies*. Cheltenham: Edward Elgar.
Karlsson, C., and B. Johansson. 2006. Dynamics and entrepreneurship in a knowledge-based economy. In *Entrepreneurship and dynamics in the knowledge economy*, ed. C. Karlsson, B. Johansson, and R.R. Stough, 12–46. New York/London: Routledge.
Karlsson, C., and R.G. Picard, ed. 2011. *Media clusters. Spatial agglomeration and content capabilities*. Cheltenham: Edward Elgar.
Karlsson, C., and P. Rouchy. 2015. Media clusters and metropolitan knowledge economy. In *Handbook on the economics of the media*, ed. R.G. Picard and S.S. Wildman, 80–106. Cheltenham: Edward Elgar.
Karlsson, C., and R.R. Stough. 2002. Introduction: Regional policy evaluation in the new economic geography. In *Regional policies and comparative advantage*, ed. B. Johansson, C. Karlsson, and R.R. Stough, 1–21. Cheltenham: Edward Elgar.
Karlsson, C., B. Johansson, and R.R. Stough, ed. 2005. *Industrial clusters and inter-firm networks*. Cheltenham: Edward Elgar.
Karlsson, C., B. Johansson, and R.R. Stough, ed. 2014. *Agglomeration, clusters and entrepreneurship. Studies in regional economic development*. Cheltenham: Edward Elgar.
Kindleberger, C.P. 1974. *The formation of financial centres: A study of comparative economic history.* Princeton, NJ: Princeton University Press.
Klepper, S. 2001. Employee startups in hugh-tech industries. *Industrial and Corporate Change 10*: 639–674.
Koopmans, T.C. 1957. *Three essays on the state of economic science*. New York: McGraw-Hill.
Krätke, S. 2003. Global media cities in a world-wide urban network. *European Planning Studies 12*: 605–628.
Krugman, P. 1980. Scale Economies, Product Differentiation and the Pattern of Trade. *American Economic Review 70*: 950–959.
Krugman, P. 1990. *Rethinking international trade*. Cambridge, MA: The MIT Press.
Krugman, P. 1991. *Geography and trade*. Cambridge, MA: The MIT Press.
Krugman, P. 1993. First nature, second nature and metropolitan location. *Journal of Regional Science 33*: 129–144.
Lakshmanan, T.R. 1989. Infrastructure and economic transformation. In *Advances in spatial theory and dynamics*, ed. Å.E. Andersson, D.F. Batten, and B. Johansson, 241–262. Amsterdam: North-Holland.
Lakshmanan, T.R., and W.G. Hansen. 1965. A retail market potential model. *Journal of the American Institute of Planners 31*: 134–143.
Lösch, A. 1943. *Die raumliche ordnung der wirtschaft*. Stuttgart: Gustav Fischer.

Lundvall, B.-Å. 2002. The learning economy: Challenges to economic theory and policy. In *A modern reader in institutional and evolutionary economics, key concepts*, ed. G.M. Hodgson, 26–47. Cheltenham: Edward Elgar.
Marshall, A. 1920. *Principles of economics*. London: Macmillan.
Martin, R., and P. Sunley. 1997. Paul krugman's geographical economics and its implications for regional theory: A critical assessment. *Regional Studies 77*: 259–292.
May, W., C. Mason, and S. Pinch. 2001. Explaining industrial agglomeration: The case of the British high-fidelity industry. *Geoforum 32*: 363–376.
McCann, P., and Z.J. Acs. 2011. Globalization: Countries, cities and multinationals. *Regional Studies 45*: 17–32.
Mills, E.S. 1967. An aggregative model of resource allocation in a metropolitan area. *American Economic Review 57*: 197–210.
Mills, E.S., and B.W. Hamilton. 1984. *Urban economics*. 3rd ed. Glenview: Scott, Foresman, and Co.
Myrdal, G. 1957. *Economic theory and under-developed regions*. London: Ducksworth.
Neff, G. 2005. The changing place of cultural production: The location of social networks in a digital media industry. *The Annals of the American Academy of Political and Social Science 597*: 134–152.
Ogawa, H. 2000. Spatial impact of information technology development. *The Annals of Regional Science 34*: 537–551.
Ohlin, B. 1933. *Interregional and international trade*. Cambridge, MA: Harvard University Press.
Pandit, N.R., G.A.S. Cook, and G.M.P. Swann. 2008. Clustering of financial services. In *Handbook of research on cluster theory*, ed. C. Karlsson, 249–260. Cheltenham: Edward Elgar.
Poon, J.P.H. 2003. Hierarchical tendencies of capital markets among international financial centres. *Growth and Change 34*: 135–156.
Porter, M.E. 1985. *Competitive advantage*. New York: Free Press.
Porter, M.E. 2000. Locations, clusters and company strategy. In *The oxford handbook of economic geography*, ed. G.L. Clark, M.P. Feldman, and M.S. Gertler, 253–274. Oxford: Oxford University Press.
Pratt, A.C. 2000. New media, the new economy and new spaces. *Geoforum 31*: 425–436.
Press, K. 2006. *A life cycle for clusters? The Dynamics of agglomeration, change, and adaptation*. Heidelberg: Physica-Verlag.
Quah, D. 1999. A weightless economy, *The Unesco courier*, Summer, 30–32
Quigley, J.M. 1998. Urban density and urban growth. *Journal of Economic Perspectives 12*: 127–138.
Richardson, G.B. 1972. The organisation of industry. *Economic Journal 82*: 883–896.
Scott, A.J. 1998. *Regions and the world economy: The coming shape of global production, competition and political order*. Oxford: Oxford University Press.
Sen, A., and T. Smith. 1995. *Gravity models of spatial interaction behaviour*. Berlin: Springer.
Slager, A. 2006. *The internationalization of banks; patterns, strategies and performance*. Basingstoke: *Palgrave Macmillan*.
Stigler, G. 1951. The division of labour is limited by the extent of the market. *Journal of Political Economy 59*: 185–193.
Stimson, R.J., R.R. Stough, and B.H. Thomas. 2002. *Regional economic development. Analysis and planning strategy*. Berlin: Springer.
Venables, A.J. 2001. *Geography and international inequalities: The impact of new technologies*, *Working paper no. 05/07*. London: London School of Economics.
von Thünen, J.H. 1826. *Der isolierte Staat in Beziehung auf nationale Ökonomie und Landwirtschaft*. Stuttgart: Gustav Fischer.
Westlund, H. 2006. *Social capital in the knowledge economy. Theory and empirics*. Berlin: Springer.

Coalitions

Myrna Wooders and Frank H. Page Jr.

Abstract

Coalitions appear in an incredible diversity of economic and game-theoretic situations, ranging from marriages, social coalitions and clubs to unions of nations. We discuss some of the major approaches to coalition theory, including models treating why and how coalitions form, equilibrium (or solution) concepts for predicting outcomes of models allowing coalition formation, and current trends in research on coalitions. We omit a number of related topics covered elsewhere in this dictionary, such as matching and bargaining.

Keywords

f-core; Abstract games; Admissible set; Asymmetric information; Bargaining; Bargaining set; Basins of attraction; Clubs; Coalitions; Cooperative games; Cores; Differential information; Domination; Epsilon core; Extensive form games; Far-sighted stability; Hedonic games; Implicit coalitions; Incomplete information; Information sharing; Inner core;

Irreversibilities; Kernel; Law of demand; Law of supply; Linear programming; Link formation; Local public goods; Myerson value; Nash equilibrium; Nash program; Network formation; Non-cooperative games; Non-transferable utility games; Owen equilibrium; Owen set; Pairwise stability; Partnered core; Private information; Public goods; Shapley value; Small group effectiveness; Solution concepts; Strong stability; Subgame perfection; Superadditivity; Supernetworks; Tau value; Tiebout hypothesis; Transferable utility games; von Neumann–Morgenstern stable set

JEL Classifications
D71; C7

The traditional notion of a coalition is a group of players who can realize some set of outcomes for its own membership. How to define this set of outcomes is a fundamental question and its definition is typically either avoided, by assuming that the set of outcomes is given, or treated simultaneously with a solution concept. Alternatively, some process may be given that plays a role in determining the set of outcomes that are achievable by each coalition.

How to define a coalition is an even more fundamental question. Typically a coalition is taken as a subset of players of a game. Yet we often perceive that individuals belong to overlapping coalitions. For example, an individual may belong to the Citizens Coalition for Responsible Media, Immunization Action Coalition and the Democratic Party. We also perceive that coalitions may be *temporary* alliances of groups of people, factions, parties, or nations. For most of this article, however, we view a coalition as simply a subset of players of a game.

When both the concepts of a coalition and its attainable set of outcomes have been defined, the question arises of how the gains from coalitional activities are to be allocated among the members of any coalition that might form, bringing us to the notion of a solution concept. A solution concept is a rule which must be satisfied by any allocation or attainable outcome that is viewed as stable or as an equilibrium. Given a description of the primitives of a situation (a game, economy, or social situation, for example) a solution concept may be viewed as predicting which outcome(s) will emerge. Implicitly, a solution concept involves assumptions about the behaviour of individuals or groups of individuals. Even in situations where a particular solution concept seems compelling, however, there may be no attainable outcomes satisfying the requirements of the solution concept. This problem, and the fact that no single solution concept seems to fit all situations, means that there are competing notions of solution concepts.

In this article we discuss issues of coalitions, the outcomes attainable by coalitions and the division of the benefits of coalition formation among the members of a coalition. Many of the fundamental questions that still intrigue researchers have their roots in the early literature of game theory. We will sketch some of the main concepts in the literature on coalitions, going back to von Neumann and Morgenstern's celebrated volume, with its notion of dominance, and also sketch some of the current approaches to questions of coalitional activities. We conclude by noting some new approaches to what a coalition might be and do and directions that research may be taking.

Domination

What a coalition can achieve, or, even more fundamentally, what a coalition can improve upon for its own membership is a fundamental question. This was realized already by von Neumann and Morgenstern (1953), who introduced the notion of domination. An imputation x (or payoff vector, listing a payoff for each participant in the society) *dominates* another imputation y with respect to a coalition S if the members of S are convinced or can be convinced that they have a positive motive for bringing about y and believe that they can do so. The coalition S is called *effective* (for x). Note that it is possible there is another payoff vector y', a coalition S' that is effective for y', and y' dominates y with respect to S' (but not with respect to S)

and in general, the relation 'dominates' may not be transitive.

Solution Concepts

A number of solution concepts based on notions of domination and effectiveness of coalitions have been defined. Three especially prominent concepts are the von Neumann–Morgenstern stable set, the Shapley value, and the core. A set V of payoff vectors, where each vector is a listing of payoffs to players in a game, is a *von Neumann–Morgenstern stable set* if (*a*) no payoff vector in V is dominated by another payoff vector in V and (*b*) every payoff vector not in V is dominated by some vector in V. The *core*, introduced in Gillies and Shapley in 1953 (see the Logistics Research Project 1957, which contains descriptions of the presentations of D. Gillies and L.S. Shapley, where the core was introduced), consists of those payoff vectors x that are feasible and undominated. The formulation of Gillies (1959) of the core of an abstract game can be widely applied. An *abstract game* consists of a set of alternatives for each coalition and a dominance relationship. The *Shapley value*, introduced in Shapley (1953), assigns to each player his *expected* marginal contribution to coalitions and is also used in numerous applications. Alternative notions of the core and of the value include the *Owen value* (Owen 1977), the τ-*value* (Tijs 1981), the *inner core* (Myerson 1995; Qin 1994; and references therein), and the *partnered core* (Albers 1979; Bennett 1983; Reny and Wooders 1996a).

Let us consider a simple example. Let $N = \{1, 2, 3\}$ be the player set. Suppose that any one player can earn zero, any two players can earn one dollar and the three players together can earn $M \geq 0$ dollars. Suppose $M = 1$; then the von Neumann–Morgenstern stable set consists of the payoff vectors $\left(\frac{1}{2}, \frac{1}{2}, 0\right)$, $\left(\frac{1}{2}, 0, \frac{1}{2}\right)$, and $\left(0, \frac{1}{2}, \frac{1}{2}\right)$. Any payoff vector (z_1, z_2, z_3) is in the core if $z_i \geq 0$ for all i A N and $z_i + z_j \geq 1$ for every pair i, j. This implies that, unless $M \geq \frac{3}{2}$, the core is empty. The Shapley value is defined for *superadditive games*, games with the property that the set of payoff vectors achievable by any union of disjoint coalitions is at least as large as the set of payoff vectors achievable by the coalitions independently.

Superadditivity, for our example, implies that $M \geq 1$, in which case the Shapley value consists of the payoff vector $(\frac{M}{3}, \frac{M}{3}, \frac{M}{3})$.

The bargaining set, introduced by Aumann and Maschler (1964), is based on threats and counter-threats. A payoff vector x is in the *bargaining set* if for every credible objection there is a credible counter-objection. That is, if there is a payoff vector y that dominates x with respect to a coalition S then there is another payoff vector y' and coalition S' that is effective for y' and y' is at least as good as x for the members of S' who are not in S and at least as good as y for members of both S and S'. There are a number of related concepts. The *kernel*, introduced in Davis and Maschler (1965), requires that objections and counter-objections have equal strengths. For our example above, the point $(\frac{M}{3}, \frac{M}{3}, \frac{M}{3})$ is also in the bargaining set and in the kernel. Recent research on concepts of the bargaining set has been spurred by the *Mas-Colell bargaining set* (Mas-Colell 1989) which adapts the bargaining set to economies with a continuum of agents and proves equivalence of the outcomes of the bargaining set and the core in an exchange economy.

Another interesting notion is the *admissible set*, introduced in Kalai and Schmeidler (1977). (See also references therein and Shenoy 1980.) Take as given a set of feasible alternatives, denoted by S, a dominance relation M and the transitive closure of M, denoted by $\widehat{M}$. The *admissible set* is the set $A(S; M) = \{x \in S : y \in S \text{ and } y \widehat{M} x \text{ imply } x \widehat{M} y\}$. The admissible set describes those outcomes that are likely to be reached by any dynamic process that respects preferences. Note that the admissible set concept can be applied to a host of game-theoretic situations, ranging from non-cooperative games, where a coalition consists of an individual player, to fully cooperative games, where any coalition can be allowed to form. As shown by Kalai and Schmeidler, under certain conditions the

admissible set coincides with the set of Nash equilibria and, for cooperative games, the admissible set coincides with the core. More recently, it has been shown that the admissible set consists of the union of basins of attraction, and a von Neumann–Morgenstern set consists of one member of each basin (Page and Wooders 2006).

Behaviour of Coalition Members

What a coalition can achieve also depends on the *behaviour of the members of the coalition*. For example, potential coalition members may bargain over the distribution of the gains to coalition formation and outcomes in the core may not be achievable as equilibria of non-cooperative bargaining processes (an important point made by John Nash 1953, leading to the *Nash program*). Chatterjee et al. (1993) demonstrate this point very well for transferable utility (TU) games, which describe what a coalition can achieve by simply a number, in interpretation, an amount of money, for example.

As stressed by Xue (1998), it may matter whether players are farsighted or myopic in their thinking about forming coalitions. Myopic players take as given the actions of others and behave accordingly. In choosing their actions, farsighted players, in contrast, take into account the reactions of other players to their actions and thus the eventual consequences of their actions. See also Diamantoudi and Xue (2003) who study the far-sighted core of a hedonic game – a game where, instead of payoff sets for coalitions, preferences are given for each individual over all coalitions in which he is contained – and Mauleon and Vannetelbosch (2004) who both allow 'spillovers' between coalitions and farsightedness of players, and demonstrate sufficient conditions for there to exist stable outcomes. (Two important papers in the game theoretic literature studying farsightedness, but not coalition formation, are Chwe 1994, and Harsanyi 1974.)

Players may also take into account 'asymmetric dependencies' within coalitions. A solution displays an asymmetric dependency if one player needs the presence of a second player to realize his payoff in the solution, but the second player does not need the presence of the first. When a player i is dependent on another player j in this sense, but j is not dependent on i, then j is in a position to attempt to obtain a larger share of the surplus from i. Consider, for example, a two-person divide-the-dollar bargaining game. Any division giving the entire dollar to one participant displays an asymmetric dependency; the player receiving the dollar is dependent on the player receiving zero. The player receiving zero is not compelled to join the two-person coalition to receive his part of the payoff. In contrast, to achieve the payoff of 50 cents for each player the two-person coalition is compelled to form – the players are partnered. The partnered core, introduced in Albers (1979) and Bennett (1983) for TU games and in Reny and Wooders (1996a) for non-transferable utility games (where the set of payoffs achievable by a coalition are described by vectors listing a payoff for each member of the coalition) consists of those outcomes in the core with the property that, to achieve his payoff, no individual needs another individual who does not need him. Even in well-behaved exchange economies there may be no outcomes in the core that are not partnered; that is, all outcomes in the core may be vulnerable to the threat of secession by some coalition of players. Page and Wooders (1996) provide an example.

Behaviour of Non-Coalition Members

In many situations, what a coalition can achieve depends on assumptions about the *behaviour of non-coalition members* (sometimes called the 'complementary coalition', although there is no requirement that the complementary coalition actually forms an alliance); for example, individuals may steal, or drop garbage in the backyards of others, or there may be widespread pollution. Two alternative definitions of the core, from Aumann and Peleg (1960), highlight the dependence of the core on the assumptions made about what outcomes are perceived as feasible by coalitions: the α-core, consisting of those outcomes that a coalition can guarantee for its membership, and the β-core, consisting of those outcomes that a coalition cannot be prevented from achieving for

its membership. In some situations, such as private goods economies without externalities or in some recent models of economies with clubs or local public goods, these two notions are equivalent, but, as noted by Shapley and Shubik (1969a), in the presence of externalities between coalitions these concepts may yield different outcomes.

Members of a coalition may also be directly affected by the structure of alliances among non-members of the coalition. This consideration underlies the Lucas and Thrall (1963) concept of *a partition function form game*, where the attainable total payoff to a coalition depend on the structure of coalitions formed by the complementary player set.

In the approach of Chander and Tulkens (1995; 1997), to predict the set of outcomes that it can achieve, a coalition presumes that the outside players will adopt their individually best reply strategies, leading to their notion of the gamma core. In the sense that the non-coalition members are treated as forming one-person coalitions, the Chander–Tulkens approach is more restrictive than that of Lucas and Thrall. When it is assumed that coalitions can freely merge or break apart and are farsighted, however, Chander (2007) demonstrates that, subsequent to a deviation by a coalition, the non-members will have incentives to break apart into singletons, thus providing a justification for the Chander–Tulkens approach.

Other approaches to the question of what a coalition can achieve for its membership have also appeared in the literature. Some recent contributions allow theft or pillage by non-coalition members; see, for example, Jordan (2006), where the payoffs attainable by a coalition are determined endogenously, and references therein.

In application, questions of the behaviour of the non-coalition members have been especially important in industrial organization and environmental economics; see, for example, Yi (1997) and Bloch (1996); see Bloch (2005) and Carraro (2005) for discussions of relevant literature.

Information Sharing Within Coalitions

When players have private information new and difficult issues arise. Chief among these is the issue of *information sharing within coalitions*. How can members of a coalition be induced to share their private information truthfully? Or, if it is not shared truthfully, how much information will be shared and how much of it will be believed? In his seminal paper, Wilson (1978) introduced two notions of the core for situations with private information, namely, the coarse core and the fine core; later Yannelis (1991) introduced the private core. Each of these core notions corresponds to assumptions about the extent to which private information of individual players is shared within coalitions. These issues are further addressed in Allen (2006), who treated core concepts in exchange economies, and Page (1997), who extended Allen's results to infinite dimensional commodity spaces. There is also the question of what informational time frame should be used in defining a solution concept. Following the informational distinctions introduced by Holmstrom and Myerson (1983) in extending the notion of Pareto efficiency to economies with private information, we can ask whether the solution concept should be *ex ante* (that is, defined relative to *ex ante* probability beliefs concerning the future information state of the economy – and therefore before players know their private information), whether it should be interim in nature (that is, defined relative to each possible profile of players' private information – and therefore after each player knows his private information but before players know the information of others), or whether it should be *ex post* (that is, defined relative to each possible information state of the economy – and therefore after each player knows the information state of the economy).

Following a mechanism design approach, Forges et al. (2002) address the issue of honest information revelation within coalitions by focusing on coalitionally incentive-compatible direct mechanisms. A coalitional direct mechanism is a mapping from the set of information profiles of coalition members into coalitional allocations. A coalitional direct mechanism is *incentive compatible* if no coalition member has an incentive to lie about his private information –on the assumption that other coalition members report their private information truthfully (that is, truthful

reporting is a Nash equilibrium of the coalitional revelation game induced by the mechanism). Formulating the coalitional mechanism design game as a TU game in characteristic function form, they demonstrate non-emptiness of the incentive compatible '*ex ante* core'. Other contributions which analyse interim core notions include Ichiishi and Idzik (1996), Hahn and Yannelis (1997), Vohra (1999), Volij (2000), Demange and Guesnerie (2001), Dutta and Vohra (2005) and Myerson (2007). See Forges et al. (2002) for a survey.

The core with incomplete information is gaining prominence in applications, such as political economy (see, for example, Serrano and Vohra 2006).

Coalition Formation

Other important questions are *how coalitions form and how coalition structures influence the behaviour of individuals within coalitions*. Several approaches are possible. Coalition formation and individual behaviour can be viewed as outcomes of market mechanisms or as outcomes of assumed cooperation within groups that may form. Alternatively, coalition formation and individual behaviour can be viewed as outcomes induced from non-cooperative behaviour. More recently coalition formation and individual behaviour within coalitions have been modelled in network settings.

The Market/Cooperative Game Approach

As suggested by Tiebout (1956) and Buchanan (1965), individuals may take as given prices for membership in coalitions (clubs, firms, jurisdictions, and so on). Tiebout conjectured that if public goods are 'local' (that is, public goods are subject to congestion and individuals can be excluded from the public goods provided in jurisdictions in which they are non-members), then the possibility of individuals moving to the jurisdictions where their wants are best satisfied subject to their budget constraints and to taxes creates a competitive 'market-like' outcome. A part of the outcome is a partition of individuals into jurisdictions. Buchanan (1965) stressed the importance of collective activities in a model of clubs with optimal club size; to illustrate, considering our example above where any two players can earn one dollar, if $M < \frac{3}{2}$, then two is the optimal club size. One way to formulate the Tiebout hypothesis (Pauly 1970; Wooders 1978; 1980) is to model the economy as one where individuals pay prices to join coalitions/clubs/jurisdictions and to demonstrate equivalence of the core and the set of outcomes of price-taking equilibrium. The results of these early papers have been greatly extended and refined; see, for example, Conley and Wooders (2001); Ellickson et al. (2001) and, for a survey, Conley and Smith (2005). The spirit of the main results is that, whenever *small group effectiveness* holds – that is, whenever all or almost all externalities can be internalized within relatively small groups of individuals (clubs, jurisdictions, firms, trading coalitions, and so on) or, in other words, whenever all or almost all gains to collective activities can be realized with some partition of the total player set into relatively small coalitions – then economies with many participants are 'market like' in the sense that price-taking economic equilibrium exists and the set of equilibrium outcomes is equivalent to the core of the economy.

The results for models of economies with local public goods and clubs suggest results for cooperative games with endogenous coalition structures. Under small group effectiveness, cooperative games with many players are 'market games' (as defined in Shapley and Shubik 1969b) and thus can be represented as economies where all individuals have concave, continuous utility functions (Wooders 1994a, b). (That the conditions of Wooders 1983, imply that games with many players are market games was first noted by Shubik and Wooders 1982, and the concavity of the limiting per capita payoff function was first explicitly noted in 1987 by Robert Aumann in his entry game theory in the first edition of this dictionary, which is reproduced in the present edition).

A simple example may provide some intuition. Suppose any two players can earn $1.00, as in our earlier example, but now suppose that there are n players in total. If n is odd, then the core is

empty, but for large n each player can receive nearly \$0.50 so certain approximate cores are non-empty and the approximation is 'close'. In defining an appropriate approximate core concept the modeller can either suppose that there are some costs to coalition formation, which can be allowed to go to zero as n instead that the payoff to a coalition with m members is a real number $v(m)$. Suppose the game is becomes large, or that a relatively small set of players can be ignored. Now, more generally, suppose essentially superadditive – the total payoff achievable by $m + m'$ players is greater than or equal to $v(m) + v(m')$. Then the only condition required to ensure non-emptiness of approximate cores of games with many players is that there is a bound K such that $\frac{v(m)}{m} \leq K$ for all m, which implies small group effectiveness. The limiting concave utility function alluded to above is $u(n) = \sup \frac{v(m)}{m} n$. See also Robert Aumann's discussion of Wooders's (1983) result in game theory.

Some other market properties of a game with many participants are that: Outcomes in the core or approximate cores treat most similar players nearly equally (Wooders 1983; Shubik and Wooders 1982; and for the most recent results, Kovalenkov and Wooders 2001a). The Shapley value is in an approximate core (Wooders and Zame 1987). A 'law of scarcity' holds; that is, increasing the abundance of one type of player leads to a decrease in the core payoffs to individual players of the same of similar types (Scotchmer and Wooders 1988; and, for recent results and references, Kovalenkov and Wooders 2005b; 2006). The law of scarcity is in the spirit of the law of demand and law of supply of private goods economies but differs in that an additional player in a game creates both creates additional demand (for the cooperation of others) and additional supply (of players of the same type).

To illustrate further the intimate relationships between markets and economies with group activities such as clubs and/or local public goods, we will discuss Owen (1975), who treats a production economy where individuals are endowed with resources that may be used in production. Rather than selling their resources to firms, individuals form coalitions and use the resources owned by the coalition to produce output which can then be sold at given prices. Owen places conditions on the model – specifically linear production functions – that ensure non-emptiness of the core of the derived game, whose coalitions consist of owners of resources. From the fundamental theorem of linear programming, associated with any point in the core of the game there is a price vector for resources, which is analogous to a competitive equilibrium price vector for resources except that the budget constraint need not be satisfied by individuals but instead only by coalitions. Owen demonstrates that, when the economy is replicated, the core converges to the set of Owen equilibrium prices. The Owen set and the Owen equilibrium prices have been studied in a number of papers – for example, Kalai and Zemel (1982), Samet and Zemel (1984), Granot (1986) and Gellekom et al. (2000). (There is also some relationship to the literature on oligopoly and cost-sharing; see, for example, Sharkey 1990; Tauman et al. 1997.)

It is easy to interpret the resources in Owen's model as attributes of individuals, such as their intelligence, skill level, wealth, ability to dance the tango, and so on. (Of course, labour is typically an input into a production process.) We can also easily interpret a coalition that forms as a club. For example, the club may be a dinner club, where each person brings himself – his personality, his gender, and so on – and also perhaps contributes a dish for the meal. The benefits to membership in a club depend on the attributes of its members – whether they are charming, whether they are good cooks. A difficulty in applying Owen's model to economies with clubs, jurisdictions, or any sort of essential group activity is that his results require linearity of the production function. However, as Owen remarks, concavity of preferences and production possibilities, as in Debreu and Scarf (1963), suffices for all his results except uniqueness of Owen equilibrium prices. But the concavity of limiting per capita payoff functions under the conditions of essential superadditivity and small group effectiveness of Wooders (1983; 1994a, b) implies that in large games with clubs or coalitional activities the economy is representable as a market economy where

individuals have concave preferences. Essential superadditivity simply allows a set of players to partition itself and achieve the outcomes achievable by the collective activities of the members of each element of the partition. Finiteness of the supremum of per capita payoffs (per capita boundedness) rules out average (per individual player) payoff from becoming infinitely large. Recent research investigates the relationship between club economies and games in more detail (see, for recent surveys, Wooders 1994b; Kovalenkov and Wooders 2005a; Conley and Smith 2005).

Closely related in important ways to the market approach are approaches that assume cooperative behaviour on the part of members of the coalitions that form. As in the market approach, what a coalition can achieve is taken as defined, a solution concept assumed (which in some cases includes a partition of the set of players into groups that can achieve their part of the outcome), and the existence and properties of outcomes satisfying the requirements of the solution concept are examined. Classic contributions to this literature, besides those mentioned above, include Aumann and Maschler (1964), Aumann and Shapley (1974), Shapley (1971), and Hart and Kurz (1983). More recent contributions include, among others, Demange (1994), Bogomolnaia and Jackson (2002), Banerjee et al. (2001), Le Breton et al. (2006), and Bogomolnaia et al. (2007). These interesting works deepen insight into the question of conditions on models ensuring there is some outcome satisfying the requirements of solutions having desirable properties, especially the core.

Necessary and sufficient conditions for non-emptiness of cores are demonstrated by Bondareva (1963) and Shapley (1967) for games with transferable utility and, most recently, by Predtetchinski and Herings (2004) and Bonnisseau and Iehle (2007) for non-transferable utility games.

A small but growing literature, initiated by the assignment games of Gale and Shapley (1962), Shapley and Shubik (1972) and Aumann and Drèze (1974), addresses the question of what conditions on permissible coalition structures will ensure that a game has a non-empty cores, independently of the sets of attainable outcomes of the game. Early papers providing such conditions are Kaneko and Wooders (1982) and Le Breton et al. (1992). Recent papers have treated sufficient conditions for non-emptiness of the core of a hedonic game, where preferences are defined directly over coalitions (Bogomolnaia and Jackson 2002; Banerjee et al. 2001; Papai 2004) while Iehle (2006) provides necessary and sufficient conditions. Demange (2004) demonstrates that imposing a hierarchical structure on the set of players, limiting the coalitions that can form, will ensure existence of an efficient outcome that is stable in the sense that no admissible coalition, called a team, could improve upon the outcome for its members. A hierarchical structure is represented by a pyramidal network. A team is a group of individuals who can communicate through the channels created by the hierarchical structure.

A related branch of literature focuses on conditions ensuring that groups of agents do not break away from a coalition. Le Breton and Weber (2001), Haimanko et al. (2004), and Drèze et al. (2006) investigate models with heterogeneous individuals and conditions ensuring existence of secession-proof outcomes, that is, outcomes that are immune to breakaways by subgroups of individuals and are thus in the core. For a different approach motivated by the idea that if a group secedes from a larger group then it does not necessarily stand alone, see Reny and Wooders (1996b), who use the solution concept of the partnered core. See also Alesina and Spolaore (1997) who demonstrate that, in a model of public good provision with a continuum of consumers who are differentiated by their preferred location for a facility and voting within each community, in equilibrium there are too many coalitions (nations).

Non-cooperative Game Approach

Coalitions can arise as equilibrium outcomes of either static or dynamic non-cooperative games. In the non-cooperative literature on clubs or local public goods, it may be assumed that there is a fixed set of jurisdictions, each providing some

level of a public good for its residents. Individuals who move to a jurisdiction pay the average cost of public good provision. Alternatively, individuals may be required to pay a proportion of their income towards financing the public good produced by the jurisdiction. Individuals each chose a jurisdiction in which to live. The main questions are whether a non-cooperative equilibrium (Nash equilibrium in pure strategies) exists and its properties, such as whether, in equilibrium, members of the same jurisdiction have similar wealths. Contributions to this literature include Greenberg and Weber (1986), Demange (1994), Konishi et al. (1997; 1998), Gravel and Thoron (2007). See also Demange (2005), who discusses literature involving both cooperative and non-cooperative approaches. Based on the concept of coalition-proofness (Bernheim et al. 1987) Conley and Konishi (2002) obtain existence of an efficient, migration-proof equilibrium for local public good (club) economies with many but a finite number of players. Casella (1992) and Casella and Feinstein (2002) consider the effects of the possibilities of trade in private goods in the formation of clubs/jurisdictions.

In a number of papers on dynamic games of coalition formation, a payoff set is given for each coalition. Suppose for simplicity that, for each coalition S, there is a unique attainable payoff vector $\{x^i(S) : i \in S\}$. If players are randomly ordered and if according to the ordering each player lists those players he would like as members of his coalition, then one possible solution to such a game of non-cooperative coalition formation would be a partition of the total player set into coalitions where for each coalition S in the partition the members of S all choose S and each player i A S receives the payoff $x^i(S)$. If player i belongs to no such coalition, then he receives some default payoff $x^i(\{i\})$. This sort of approach was introduced in Selten (1981). Perry and Reny (1994) provide a non-cooperative implementation of the core for TU games. In the Perry–Reny model proposed, time is continuous. This ensures that there is always time to reject a non-core proposal before it is consummated. Which coalitions will form typically depends crucially on the rules of the game. The Perry–Reny implementation is meant to reflect the standard motivation for the core as closely as possible. Hart and Mas-Colell (1996) implement the *consistent value* (Maschler and Owen 1992) for NTU games, which, for TU games, is equivalent to the Shapley value. Bloch (1996) treats games where, as in the Lucas–Thrall model, the payoff achievable by a group of players may depend on the entire coalition structure of the remaining players. Ray and Vohra (1997; 1999) study coalitional agreements and coalitional bargaining in partition function games. See Bandyopadhyay and Chatterjee (2006) for a survey of coalition formation based on non-cooperative bargaining. See also Myerson (1995), Seidmann and Winter (1998), Mauleon and Vannetelbosch (2004), among others.

Networks and Coalition Formation

Because networks allow for a detailed specification of interactions between individuals and between coalitions, abstract games over networks have a greater potential to capture the subtleties of bargaining and negotiation than do the abstract coalitional form games of von Neumann–Morgenstern and Gillies and Shapley. A seminal contribution to this line of research is the paper by Myerson (1977). Myerson begins by assuming that the worth of each possible coalition depends on the structure of cooperation between individuals as given by a graph where nodes represent individuals and links between nodes represent interactions between individuals. As in much of the subsequent literature Myerson imposes an allocation rule, a rule specifying how the worth of a coalition is to be shared among its members. The worth of any connected (linked) set of players is divided according to the rule. The specific rule chosen by Myerson is a variant of the Shapley value, now known as the Myerson value. As Myerson shows, this is the only rule satisfying both component efficiency (in sum, the members of each component of the network receive the worth of that component as a coalition) and a fairness property that requires any two players to benefit equally from the formation of a link. Aumann and Myerson (1988) work with extensive form games, where players choose links strategically and allow players to look ahead

and to take into account the end effects of their actions. In their model, once a link is formed, it cannot be broken. The equilibrium concept is non-cooperative subgame perfection. Once players have formed links, the payoffs to players are determined by the Myerson value.

Jackson and Wolinsky (1996) also treat link formation between individual players. A network satisfies their pairwise stability condition if no two players could benefit by creating a link between them and no one player could benefit by cutting a link with another player. Based on the Jackson–Wolinsky model, numerous papers have now looked at costs and benefits of link formation between players and equilibrium outcomes; see Dutta, van den Nouweland, and Tijs (1998) for example, and van den Nouweland (2005) for some recent results and a review. Herings et al. (2006) introduce notions of pairwise farsighted stability. Jackson and van den Nouweland (2005) introduce the concept of a strongly stable network. A network is strongly stable if no coalition could benefit by making changes (additions or deletions) to the links of coalition members. As Jackson and van den Nouweland show, the existence of strongly stable networks is equivalent to non-emptiness of the core in a derived cooperative game. See also Jackson and Watts (2002), who use linking networks and stochastic dynamics to study the evolution of networks.

Other recent works addressing questions of coalition formation in networks make assumptions concerning what a coalition believes it can achieve. These contributions include Watts (2001), who assumes that dominance must be direct, in the sense that a coalition will act to change a network from g to g' only if it perceives an immediate gain. In contrast, Page et al. (2005) consider indirect dominance where a network g dominates another network g' if there is a coalition S that believes it can trigger a series of changes beginning with the network g and ending with the network g' that is preferred by all members of S. Whether dominance is direct or indirect is of crucial importance, as illustrated in Diamantoudi and Xue (2003) and Page and Wooders (2007), among others. Consider, for example, a situation with two jurisdictions, say J_1 and J_2 and seven people. Each person would like to live in the jurisdiction with the fewest residents. With direct dominance, any partition of the people between the two jurisdictions with three people in one jurisdiction and four in the other is stable. In contrast, with indirect dominance, the situation changes; players can be more optimistic. Suppose that initially there are four people in jurisdiction J_1 and three in J_2. Two people in J_1 may move into J_2 in the belief that, since J_2 has become so crowded, three people will leave J_2 and move to J_1, with the result that the two initial movers will be better off.

Using supernetworks, introduced in Page et al. (2005), where nodes represent networks and directed arcs represent coalitional moves and coalitional preferences, networks can also provide a simple representation of the rules of network formation and hence the rules of coalition formation. Network formation rules play a crucial role in determining coalitional outcomes. To illustrate, in the literature on markets and on cooperative games, it is assumed that coalitions can exclude individuals. It may be, however, that groups (or coalitions) are subject to 'free entry' – any group of players can freely join another group without the consent of those being joined. This has long been important in the literature on economies with clubs/local public goods; compare, for example, the models of Konishi et al. (1998) and Demange (1994) with that of Conley and Wooders (2001). As a special case, networks can also accommodate a systematic analysis of coalition formation and payoff division when there are potential irreversibilities. For example, given the informational environment, it may be that the only coalitions which can form are sub-coalitions of existing coalitions. Or the rules of network formation may not allow cycles.

How to Define a Coalition

The traditional approach of cooperative game theory models a coalition as an alliance of players who take as given a well-defined set of possible outcomes or payoffs. The alliance, when considering whether to 'block' a proposed outcome, is

faced with the alternative of standing alone. In reality, however, we observe that individuals belong to multiple, possibly overlapping alliances. This fact has received remarkably little attention in the literature. Some papers in the club literature allow individuals to belong to multiple clubs for the purposes of local public good provision and private good production within each club, including Shubik and Wooders (1982), Ellickson et al. (2001) and Allouch and Wooders (2006). Roughly, if there is only a finite set of sorts of clubs, bounded in size, (Ellickson et al.) or if 'per capita payoffs' are bounded (Allouch and Wooders), then in large economies the core and the set of price taking equilibrium outcomes are equivalent. An interesting application of the idea of overlapping coalitions is developed in Conconi and Perroni (2002), who assume that a country can enter into different alliances, where each alliance to which it belongs is concerned with a different issue.

The definition of a coalition also becomes an issue when the total player set is an atomless continuum. There are two approaches. One approach, introduced in Aumann (1964), is to model a coalition as a subset of positive measure. Major theorems using this approach and relating to coalitions demonstrate equivalence of the core and outcomes of price-taking equilibrium of models of economies Another approach is to describe a coalition as a finite set of players, as in Keiding (1976). This has the advantage that individuals may interact with other individuals, and permits matching or marriage models, for example. An obvious difficulty with such an approach is that, at the heart of economics, is the problem of relative scarcities. Think of the diamond–water paradox; even though water is essential for life itself, it is abundant and thus inexpensive, while diamonds are relatively inessential but scarce and thus expensive.

To see the difficulty in retaining relative scarcities while allowing finite coalitions, suppose, for example, that the points in the interval [0,2] represent boys and the points in the interval [3,4] represent girls so that there are 'twice' as many boys as girls. Suppose the only effective coalitions consist of either boy, girls pairs (i, j) where $i \in [0; 2]$ and $j \in =3; 4]$, or singletons – a matching model. Consider the set of coalitions $(i; j) : j = 3 + \frac{1}{2}i$; this set describes a partition of the total player set and marries each boy to a girl; clearly this partition is not consistent with the relative scarcities given by Lebesgue measure. Indeed, since there are one-to-one mappings of a set of positive measure onto a set of measure zero, it is even possible to have partitions of the total player set into boy–girl pairs and singletons that match each boy to a girl while leaving a set of girls of measure 1 unmatched! A solution to this problem was proposed by Kaneko and Wooders (1986) with the introduction of *measurement-consistent* partitions. A simple formulation of measurement consistency has recently been provided (Allouch et al. 2006), and we use it here. Define an *index set* for a partition of a continuum of players as one member from each element of the partition. A partition of players into finite coalitions is 'measurement-consistent' if every index set for the partition has the same measure. The partition given above is not measurement-consistent while the partition $\{(i; j) : j = 3\ )\ i; i \in [0, 1] \cup \{\{i\} : i \in (1; 2]\}\}$ is measurement-consistent. While in models of exchange economies, the core with finite coalitions (the *f*-core) and the Aumann core yield equivalent outcomes, in the presence of widespread externalities, such as global pollution, the *f*-core coincides with the set of competitive equilibrium prices while the Aumann core may be empty and, even if non-empty, may have an empty intersection with the set of equilibrium outcomes; the concepts of the Aumann core and the *f*-core are distinct with the *f*-core apparently most closely related to the set of competitive equilibrium prices (Kaneko and Wooders 1986; Hammond, Kaneko and Wooders 1989; Kaneko and Wooders 1994). Other works using the ore approach include Berliant and Edwards (2004) and Legros and Newman (1996, 2002). These papers illustrate the advantage of the *f*-core approach in that it enables analysis of activities within groups (firms or clubs, or other organizations) that may contain any finite number of individuals but are negligible relative to the entire economy.

An interesting difference between the Aumann-core and the *f*-core is that, while the Aumann-core has been axiomatized by Dubey and Neyman (1984), the authors stress that the axiomatization is completely different than axiomatizations for the core in cooperative games with only a finite number of players. In contrast, Winter and Wooders (1994) provide an axiomatization for the core of a game with finite coalitions that applies whether the player set is finite or an atomless continuum.

Conclusions

This article began with some of the first works on coalitions in the literature of game theory and concluded with recent work on coalitions and networks. It becomes apparent that the concepts of early works underlie much of even the most recent research. We see at least a part of the future of coalition theory in network modelling of socio-economic coalitions and in more behavioural approaches to coalition theory, involving 'implicit' and 'tacit' coalitions. Language and the ability to communicate well are clearly involved; see multilingualism and references there. Instead of being bound together by commitments and contracts, members of an implicit coalition may be bound together by common language, culture, objectives or by common group memberships and, even though there may be no explicit agreement, members of an implicit coalition might act together, as if they were a coalition. This raises questions of to what extent individuals, who share common group memberships as in Durlauf (2002) for example, are an implicit coalition and whether such individuals have tendencies to form more explicit coalitions. While much has been done on coalitions, there remains much to do.

See Also

- ▶ Bargaining
- ▶ Core convergence
- ▶ Game theory
- ▶ Multilingualism
- ▶ Network formation

Bibliography

Agastya, M. 1999. Perturbed adaptive dynamics in coalition form games. *Journal of Economic Theory* 89: 207–233.

Albers, W. 1979. Core-and-kernel variants based on imputations and demand profiles. In *Game theory and related topics*, ed. O. Moeschlin and D. Pallaschke. Amsterdam: North-Holland.

Alesina, A., and E. Spolaore. 1997. On the number and size of nations. *Quarterly Journal of Economics* 112: 1027–1056.

Allen, B. 1992. *Incentives in market games with asymmetric information: The core*, CORE Discussion Paper No. 9221, Université Catholique de Louvain.

Allen, B. 2006. Market games with asymmetric information: The core. *Economic Theory* 29: 465–487.

Allouch, N., J. Conley, and M. Wooders. 2006. *Anonymous price taking equilibrium in Tiebout economies with a continuum of agents; existence and characterization*, Working paper, Department of Economics, Vanderbilt University.

Allouch, N., and M. Wooders. 2006. *Price taking equilibrium in club economies with multiple memberships and unbounded club sizes*, Working paper, Department of Economics, Vanderbilt University.

Aumann, R.J. 1964. Markets with a continuum of traders. *Econometrica* 32: 39–50.

Aumann, R.J., and J. Drèze. 1974. Cooperative games with coalition structures. *International Journal of Game Theory* 3: 217–237.

Aumann, R.J., and M. Maschler. 1964. The bargaining set for cooperative games. In *Advances in game theory, annals of mathematics study 52*, ed. M. Dresher, L.S. Shapley, and A.W. Tucker. Princeton: Princeton University Press.

Aumann, R.J., and R. Myerson. 1988. Endogenous formation of links between players and of coalitions: An application of the Shapley value. In *The Shapley value*, ed. A.E. Roth. Cambridge: Cambridge University Press.

Aumann, R.J., and B. Peleg. 1960. Von Neumann–Morgenstern solutions to cooperative games without side payments. *Bulletin of the American Mathematical Society* 66: 173–179.

Aumann, R.J., and L.S. Shapley. 1974. *Values of non-atomic games*. Princeton: Princeton University Press.

Bag, P., and E. Winter. 1999. Simple subscription mechanisms for excludable public goods. *Journal of Economic Theory* 87: 72–94.

We are grateful to Harold Kuhn for his generous assistance in tracking down the origins of the concept of the core.

Bandyopadhyay, S., and K. Chatterjee. 2006. Coalition theory and applications: A survey. *Economic Journal* 116: 136–156.

Banerjee, S., H. Konishi, and T. Sonmez. 2001. Core in a simple coalition formation game. *Social Choice and Welfare* 18: 135–158.

Barberà, S., and A. Gerber. 2003. On coalition formation: Durable coalition structures. *Mathematical Social Sciences* 45: 185–203.

Bennett, E. 1983. The aspiration approach to predicting coalition formation in side payments games. *International Journal of Game Theory* 12: 1–28.

Berliant, M., and J.H.Y. Edwards. 2004. Efficient allocations in club economies. *Journal of Public Economic Theory* 6: 43–63.

Bernheim, D., B. Peleg, and M. Whinston. 1987. Coalition proof Nash equilibria I: Concepts. *Journal of Economic Theory* 42: 1–12.

Bloch, F. 1995. Endogenous structures of association in oligopolies. *RAND Journal of Economics* 26: 537–556.

Bloch, F. 1996. Sequential formation of coalitions in games with fixed payoff division and externalities. *Games and Economic Behavior* 14: 90–123.

Bloch, F. 2005. Group and network formation in industrial organization: A survey. In *Group formation in economics; networks, clubs and coalitions*, ed. G. Demange and M. Wooders. Cambridge: Cambridge University Press.

Bogomolnaia, A., and M.O. Jackson. 2002. The stability of hedonic coalition structures. *Games and Economic Behavior* 38: 201–230.

Bogomolnaia, A., M. Le Breton, A. Savvateev, and S. Weber. 2007. Stability of jurisdiction structures under the equal share and median rules. *Economic Theory.*

Bondareva, O. 1963. Some applications of linear programming to the theory of cooperative games [in Russian]. *Problemy kibernetiki* 10. English translation by K. Takeuchi and E. Wesley in *Selected Russian Papers on Game Theory 1959–1965*, ed. L. Billera, D. Cohen, and R. Cornwall. Princeton: Princeton University Press, 1968.

Bonnisseau, J.-M., and V. Iehle. 2007. Payoff-dependent balancedness and core. *Games and Economic Behavior* 61: 1–26.

Bossert, W., and Y. Sprumont. 2002. Core rationalizability in two-agent exchange economies. *Economic Theory* 20: 777–791.

Buchanan, J.M. 1965. An economic theory of clubs. *Economica* 33: 1–14.

Burani, N., and W. Zwicker. 2003. Coalition formation games with separable preferences. *Mathematical Social Sciences* 45: 27–52.

Carraro, C. 2005. Institution design for managing global commons: Lessons from coalition theory. In *Group formation in economics: Networks, clubs and coalitions*, ed. G. Demange and M. Wooders. Cambridge: Cambridge University Press.

Casella, A. 1992. On markets and clubs: Economic and political integration of regions with unequal productivity. *American Economic Review* 82: 115–121.

Casella, A., and J. Feinstein. 2002. Public goods in trade: On the formation of markets and jurisdictions. *International Economic Review* 43: 437–462.

Chander, P. 2007. The gamma-core and coalition formation. *International Journal of Game Theory* (forthcoming).

Chander, P., and H. Tulkens. 1995. A core-theoretic solution for the design of cooperative agreements on transfrontier pollution. *International Tax and Public Finance* 2: 279–293.

Chander, P., and H. Tulkens. 1997. The core of an economy with multilateral environmental externalities. *International Journal of Game Theory* 26: 379–401.

Chatterjee, K., B. Dutta, D. Ray, and K. Sengupta. 1993. A non-cooperative theory of coalitional bargaining. *Review of Economic Studies* 60: 463–477.

Chwe, M.S.-Y. 1994. Farsighted coalitional stability. *Journal of Economic Theory* 63: 299–325.

Conconi, P., and C. Perroni. 2002. Issue linkage and issue tie-in in multilateral negotiations. *Journal of International Economics* 57: 423–447.

Conley, J.P., and H. Konishi. 2002. Migration-proof Tiebout equilibrium: Existence and asymptotic efficiency. *Journal of Public Economics* 86: 243–262.

Conley, J., and S. Smith. 2005. Coalitions and clubs: Tiebout equilibrium in large economies. In *Group formation in economics: Networks, clubs and coalitions*, ed. G. Demange and M. Wooders. Cambridge: Cambridge University Press.

Conley, J., and M. Wooders. 2001. Tiebout economics with differential genetic types and endogenously chosen crowding characteristics. *Journal of Economic Theory* 98: 261–294.

Davis, M., and M. Maschler. 1963. Existence of stable payoff configurations for cooperative games. *Bulletin of the American Mathematical Society* 69: 106–108.

Davis, M., and M. Maschler. 1965. The kernel of a cooperative game. *Naval Research Logistics Quarterly* 12: 223–259.

Debreu, G., and H. Scarf. 1963. A limit theorem on the core of an economy. *International Economic Review* 4: 235–246.

De Clippel, G., and E. Minelli. 2005. Two remarks on the inner core. *Games and Economic Behavior* 50: 143–154.

Demange, G. 1994. Intermediate preferences and stable coalition structures. *Journal of Mathematical Economics* 23: 45–48.

Demange, G. 2004. On group stability in hierarchies and networks. *Journal of Political Economy* 112: 754–778.

Demange, G. 2005. The interaction of increasing returns and preferences diversity. In *Group formation in economics: Networks, clubs and coalitions*, ed. G. Demange and M. Wooders. Cambridge: Cambridge University Press.

Demange, G., and R. Guesnerie. 2001. On coalitional stability of anonymous interim mechanisms. *Economic Theory* 18: 367–389.

Diamantoudi, E., and L. Xue. 2003. Farsighted stability in hedonic games. *Social Choice and Welfare* 21: 39–61.

Drèze, J., M. Le Breton, and S. Weber. 2006. Rawlsian pricing of access to public facilities: A unidimensional illustration. *Journal of Economic Theory* 136: 759–766.

Dubey, P., and A. Neyman. 1984. Payoffs in nonatomic economies: An axiomatic approach. *Econometrica* 52: 1129–1150.

Durlauf, S. 2002. The memberships theory of poverty: The role of group affiliations in determining socioeconomic outcomes. In *Understanding Poverty in America*, ed. S. Danziger and R. Haveman. Cambridge, MA: Harvard University Press.

Dutta, B., A. van den Nouweland, and S. Tijs. 1998. Link formation in cooperative situations. *International Journal of Game Theory* 27: 245–256.

Dutta, B., and R. Vohra. 2005. Incomplete information, credibility and the core. *Mathematical Social Sciences* 50: 148–165.

Echenique, F., and J. Oviedo. 2004. Core many-to-one matchings by fixed-point methods. *Journal of Economic Theory* 115: 358–376.

Economides, N.S. 1986. *Non-cooperative equilibrium coalition structures*, University Discussion Paper No. 273, Columbia University.

Einy, E., D. Moreno, and D. Monderer. 1999. On the least core and the Mas-Colell bargaining set. *Games and Economic Behavior* 28: 181–188.

Einy, E., D. Moreno, and B. Shitovitz. 2000. On the core of an economy with differential information. *Journal of Economic Theory* 94: 262–270.

Ellickson, B., B. Grodal, S. Scotchmer, and W. Zame. 2001. Clubs and the market: Large finite economies. *Journal of Economic Theory* 101: 40–77.

Engelbrecht-Wiggans, R., and D. Granot. 1985. On market prices in linear production games. *Mathematical Programming* 32: 366–370.

Epstein, L., and M. Marinacci. 2001. The core of large differentiable TU games. *Journal of Economic Theory* 100: 235–273.

Farrell, J., and S. Scotchmer. 1988. Partnerships. *Quarterly Journal of Economics* 103: 279–297.

Forges, F., J.-F. Mertens, and R. Vohra. 2002a. The ex ante incentive compatible core in the absence of wealth effects. *Econometrica* 70: 1865–1892.

Forges, F., E. Minelli, and R. Vohra. 2002b. Incentives and the core of an exchange economy: A survey. *Journal of Mathematical Economics* 38: 1–41.

Gale, D., and L.S. Shapley. 1962. College admissions and the stability of marriage. *American Mathematical Monthly* 69: 9–15.

Garratt, R., and C.-Z. Qin. 1997. On a market for coalitions with indivisible agents and lotteries. *Journal of Economic Theory* 77: 81–101.

Gellekom, J.R.G., J.A.M. Potters, J.H. Reijnierse, M.C. Engel, and S.H. Tijs. 2000. Characterization of the Owen set of linear production processes. *Games and Economic Behaviour* 32: 139–156.

Gillies, D.B. 1959. Solutions to general non-zero-sum games. In *Contributions to the theory of games*, vol. 4, ed. A.W. Tucker and R.D. Luce. Princeton: Princeton University Press.

Granot, D. 1986. A generalized linear production model: A unifying model. *Mathematical Programming* 34: 212–222.

Gravel, N., and S. Thoron. 2007. Does endogenous formation of jurisdictions lead to wealth-stratification? *Journal of Economic Theory* 132: 569–583.

Greenberg, J. 1995. Coalition structures. In *Handbook of game theory with economic applications*, ed. R.-J. Aumann and S. Hart. Amsterdam: North-Holland.

Greenberg, J., and S. Weber. 1986. Strong Tiebout equilibrium under restricted preferences domain. *Journal of Economic Theory* 38: 101–117.

Greenberg, J., and S. Weber. 1993. Stable coalition structures with a unidimensional set of alternatives. *Journal of Economic Theory* 60: 62–82.

Hahn, G., and N. Yannelis. 1997. Efficiency and incentive compatibility in differential information economies. *Economic Theory* 10: 383–411.

Haimanko, O., M. Le Breton, and S. Weber. 2004. Voluntary formation of communities for the provision of public projects. *Journal of Economic Theory* 115: 1–34.

Hammond, P., M. Kaneko, and M. Wooders. 1989. Continuum economies with finite coalitions: Core, equilibria, and widespread externalities. *Journal of Economic Theory* 49: 113–134.

Harsanyi, J. 1974. An equilibrium point interpretation of stable sets and a proposed alternative definition. *Management Science* 20: 1472–1495.

Hart, S., and M. Kurz. 1983. Endogenous formation of coalitions. *Econometrica* 51: 1047–1064.

Hart, S., and A. Mas-Colell. 1996. Bargaining and value. *Econometrica* 64: 357–380.

Herings, J-J., A. Mauleon, and V. Vannetelbosch. 2006. *Farsightedly stable networks*. RM/06/041, University of Maastricht.

Herings, J.-J., G. van der Laan, and D. Talman. 2007. The socially stable core in structured transferable utility games. *Games and Economic Behavior* 59: 85–104.

Holmstrom, B., and R. Myerson. 1983. Efficient and durable decision rules with incomplete information. *Econometrica* 51: 1799–1819.

Ichiishi, T., and A. Idzik. 1996. Bayesian cooperative choice strategies. *International Journal of Game Theory* 25: 455–473.

Iehle, V. 2006. *The core partition of a hedonic game*, Working paper, Maison des Sciences Economiques, Université Panthéon-Sorbonne.

Izquierdo, J.M., and C. Rafels. 2001. Average monotonic cooperative games. *Games and Economic Behavior* 36: 174–192.

Jackson, M., and A. van den Nouweland. 2005. Strongly stable networks. *Games and Economic Behavior* 51: 420–444.

C

Jackson, A., and A. Watts. 2002. The evolution of social and economic networks. *Journal of Economic Theory* 106: 265–295.

Jackson, A., and A. Wolinsky. 1996. A strategic model of social and economic networks. *Journal of Economic Theory* 71: 44–74.

Jordan, J.S. 2006. Pillage and property. *Journal of Economic Theory* 131: 26–44.

Kajii, A. 1992. A generalization of Scarf's theorem: An α-core existence theorem without transitivity or completeness. *Journal of Economic Theory* 56: 194–205.

Kalai, E., A. Postlewaite, and J. Roberts. 1979. A group incentive compatible mechanism yielding core allocations. *Journal of Economic Theory* 20: 13–22.

Kalai, E., and D. Schmeidler. 1977. An admissible set occurring in various bargaining situations. *Journal of Economic Theory* 14: 402–411.

Kalai, E., and E. Zemel. 1982. Generalized network problems yielding totally balanced games. *Operations Research* 30: 998–1008.

Kaneko, M., and M. Wooders. 1982. Cores of partitioning games. *Mathematical Social Sciences* 3: 313–327.

Kaneko, M., and M. Wooders. 1986. The core of a game with a continuum of players and finite coalitions: The model and some results. *Mathematical Social Sciences* 12: 105–137.

Kaneko, M., and M. Wooders. 1989. The core of a continuum economy with widespread externalities and finite coalitions: From finite to continuum economics. *Journal of Economic Theory* 49: 135–168.

Kaneko, M., and M. Wooders. 1994. Widespread externalities and perfectly competitive markets. In *Imperfections and behavior in economic organizations*, ed. R. Gilles and P. Ruys. Boston: Kluwer Academic Publishers.

Kaneko, M., and M. Wooders. 1996. The nonemptiness of the *f*-core of a game without side payments. *International Journal of Game Theory* 25: 245–258.

Keiding, H. 1976. Cores and equilibria in an infinite economy. In *Computing equilibrium: How and why*, ed. J. Los and M.W. Los. Amsterdam: North-Holland.

Kóczy, Á., and L. Lauwers. 2004. The coalition structure core is accessible. *Games and Economic Behavior* 48: 86–93.

Konishi, H., M. Le Breton, and S. Weber. 1997. Equilibrium in a model with partial rivalry. *Journal of Economic Theory* 72: 225–237.

Konishi, H., M. Le Breton, and S. Weber. 1998. Equilibrium in a finite local public goods economy. *Journal of Economic Theory* 79: 224–244.

Konishi, H., and D. Ray. 2003. Coalition formation as a dynamic process. *Journal of Economic Theory* 110: 1–41.

Kovalenkov, A., and M. Wooders. 2001a. Epsilon cores of games with limited side payments; nonemptiness and equal treatment. *Games and Economic Behavior* 36: 193–218.

Kovalenkov, A., and M. Wooders. 2001b. An exact bound on epsilon for non-emptiness of the epsilon-cores of games. *Mathematics of Operations Research* 26: 654–678.

Kovalenkov, A., and M. Wooders. 2003. Approximate cores of games and economies with clubs. *Journal of Economic Theory* 110: 87–120.

Kovalenkov, A., and M. Wooders. 2005a. Many sided matchings, clubs and market games. In *Group formation in economics; networks, clubs and coalitions*, ed. G. Demange and M. Wooders. Cambridge: Cambridge University Press.

Kovalenkov, A., and M. Wooders. 2005b. Laws of scarcity for a finite game – Exact bounds on estimations. *Economic Theory* 26: 383–396.

Kovalenkov, A., and M. Wooders. 2006. Comparative statics and laws of scarcity for games. In *Rationality and equilibrium; a symposium in honor of Marcel K. Richter*, ed. C.D. Aliprantis et al. Berlin: Springer-Verlag.

Kurz, M. 1989. Game theory and public economics. In *Handbook of game theory*, vol. 2, ed. R.J. Aumann and S. Hart. Amsterdam: North-Holland.

Le Breton, M., I. Ortuno-Ortin, and S. Weber. 2006. *Gamson's Law and hedonic games*, Working Paper No. 420, IDEI, Toulouse.

Le Breton, M., G. Owen, and S. Weber. 1992. Strongly balanced cooperative games. *International Journal of Game Theory* 20: 419–427.

Le Breton, M., and S. Weber 2001. *The art of making everybody happy: How to prevent a secession*, Working Paper No. 01/176, International Monetary Fund.

Lee, D., and O. Volij. 2000. The core of economies with asymmetric information: An axiomatic approach. *Journal of Mathematical Economics* 38: 43–63.

Legros, P., and A.F. Newman. 2002. Monotone matching in perfect and imperfect worlds. *Review of Economic Studies* 69: 925–942.

Legros, P., and A.F. Newman. 1996. Wealth effects, distribution, and the theory of organization. *Journal of Economic Theory* 70: 312–341.

Logistics Research Project. 1957. Reports of three informal conferences on the theory of games. Department of Mathematics, Princeton University. Accession No. (OCoLC)ocm39726894. Conferences held 20–1 March, 1953; 31 January–1 February 1955, and 11–12 March, 1957, Princeton.

Lucas, W., and R. Thrall. 1963. n-person games in partition function form. *Naval Research Logistics Quarterly* 10: 281–298.

Maschler, M., and G. Owen. 1992. The consistent Shapley value for games without side payments. In *Rational interaction*, ed. R. Selten. Berlin: Springer-Verlag.

Maschler, M., and B. Peleg. 1966. A characterization, existence proof and dimensions bounds for the kernel of a game. *Pacific Journal of Mathematics* 18: 289–328.

Maschler, M., and B. Peleg. 1967. The structure of the kernel of a cooperative game. *SIAM Journal of Applied Mathematics* 15: 569–604.

Maschler, M., B. Peleg, and L.S. Shapley. 1971. The kernel and bargaining set for convex games. *International Journal of Game Theory* 1: 73–93.

Mas-Colell, A. 1989. An equivalence theorem for a bargaining set. *Journal of Mathematical Economics* 18: 129–139.

Mauleon, A., and V.J. Vannetelbosch. 2004. Farsightedness and cautiousness in coalition formation games with positive spillovers. *Theory and Decision* 56: 291–324.

McLean, R., and A. Postlewaite. 2005. Core convergence with asymmetric information. *Games and Economic Behavior* 50: 58–78.

Milchtaich, I. 1996. Congestion games with player-specific payoff functions. *Games and Economic Behavior* 13: 111–124.

Milchtaich, I., and E. Winter. 2002. Stability and segregation in group formation. *Games and Economic Behavior* 38: 318–346.

Milleron, J.C. 1972. Theory of value with public goods: A survey article. *Journal of Economic Theory* 5: 419–477.

Monderer, D., and L.S. Shapley. 1996. Potential games. *Games and Economic Behavior* 14: 124–143.

Muench, T. 1972. The core and the Lindahl equlibrium of an economy with public goods; an example. *Journal of Economic Theory* 4: 241–255.

Myerson, R.B. 1977. Graphs and cooperation in games. *Mathematics of Operations Research* 2: 225–229.

Myerson, R.B. 1995. Sustainable matching plans with adverse selection. *Games and Economic Behavior* 9: 35–65.

Myerson, R.B. 2007. Virtual utility and the core for games with incomplete information. *Journal of Economic Theory* 136: 260–285.

Nash, J. 1953. Two-person cooperative games. *Econometrica* 21: 128–140.

Owen, G. 1975. On the core of linear production games. *Mathematical Programming* 9: 358–370.

Owen, G. 1977. Values of games with a priori unions. In *Essays in mathematical economics and game theory*, ed. R. Henn and O. Moeschlin. Berlin/Heidelberg/New York: Springer.

Page Jr., F.H. 1997. Market games with differential information and infinite dimensional commodity spaces: The core. *Economic Theory* 9: 151–159.

Page Jr., F.H., and M. Wooders. 1996. The partnered core and the partnered competitive equilibrium. *Economics Letters* 52: 143–152.

Page, F.H. Jr., and M. Wooders. 2006. *Strategic basins of attraction, the path dominance core, and network formation games*, Working Paper No. 06-W14 (revised), Department of Economics, Vanderbilt University.

Page, F.H. Jr., and M. Wooders. 2007. Networks and clubs. *Journal of Economic Behavior & Organization* (forthcoming).

Page Jr., F.H., M. Wooders, and S. Kamat. 2005. Networks and farsighted stability. *Journal of Economic Theory* 120: 257–269.

Papai, S. 2004. Unique stability in simple coalition formation games. *Games and Economic Behavior* 48: 337–354.

Pauly, M. 1970. Cores and clubs. *Public Choice* 9: 53–65.

Peleg, B. 1985. The axiomatization of the core of cooperative games without side payments. *Journal of Mathematical Economics* 14: 203–214.

Perry, M., and P.J. Reny. 1994. A noncooperative view of coalition formation and the core. *Econometrica* 62: 795–817.

Predtetchinski, A., and J.-J. Herings. 2004. A necessary and sufficient condition for non-emptiness of the core of a non-transferable utility game. *Journal of Economic Theory* 116: 84–92.

Qin, C.-Z. 1994. The inner core of an n-person game. *Games and Economic Behavior* 6: 431–444.

Ray, D., and R. Vohra. 1997. Equilibrium binding agreements. *Journal of Economic Theory* 73: 30–78.

Ray, D., and R. Vohra. 1999. A theory of endogenous coalition structures. *Games and Economic Behavior* 26: 286–336.

Reny, P.J., and M. Wooders. 1996a. The partnered core of a game without side payments. *Journal of Economic Theory* 70: 298–311.

Reny, P.J., and M. Wooders. 1996b. Credible threats of secession, partnership, and commonwealths. In *Understanding strategic interaction: Essays in honor of Reinhard Selten*, ed. W. Albers et al. Berlin: Springer-Verlag.

Roth, A.E. 1984. Stable coalition formation: Aspects of a dynamic theory. In *Coalitions and collective action*, ed. M. Holler. Würzberg: Physica-Verlag.

Samet, D., and E. Zemel. 1984. On the core and dual set of linear programming games. *Mathematics of Operations Research* 9: 309–316.

Scotchmer, S., and M. Wooders. 1988. *Monotonicity in games that exhaust gains to scale*, Technical Report No. 525, IMSSS, Stanford University.

Seidmann, D.J., and E. Winter. 1998. Gradual coalition formation. *Review of Economic Studies* 65: 793–815.

Serrano, R., and R. Vohra. 2006. Information transmission in coalitional voting games. *Journal of Economic Theory* (forthcoming).

Selten, R. 1981. A non-cooperative model of characteristic function bargaining. In *Essays in game theory and mathematical economics in honor of O. Morgenstern*, ed. V. Böhm and H. Nachtkamp. Mannheim: Bibliographisches Institut.

Shapley, L.S. 1953. A value for n-person games. In *Contributions to the theory of games II*, ed. A.W. Tucker and R.D. Luce. Princeton: Princeton University Press.

Shapley, L.S. 1967. On balanced sets and cores. *Navel Research Logistics Quarterly* 9: 45–48.

Shapley, L.S. 1971. Cores of convex games. *International Journal of Game Theory* 1: 11–26.

Shapley, L.S., and M. Shubik. 1966. Quasi-cores in a monetary economy with nonconvex preferences. *Econometrica* 34: 805–827.

Shapley, L.S., and M. Shubik. 1969a. On the core of an economic system with externalities. *American Economic Review* 59: 678–684.
Shapley, L.S., and M. Shubik. 1969b. On market games. *Journal of Economic Theory* 1: 9–25.
Shapley, L.S., and M. Shubik. 1972. The assignment game 1: The core. *International Journal of Game Theory* 1: 11–30.
Shapley, L.S., and M. Shubik. 1975. Competitive outcomes in the cores of market games. *International Journal of Game Theory* 4: 229–237.
Sharkey, W.W. 1990. Cores of games with fixed costs and shared facilities. *International Economic Review* 31: 245–262.
Shenoy, P.P. 1979. On coalition formation: A game-theoretic approach. *International Journal of Game Theory* 8: 133–164.
Shenoy, P.P. 1980. A dynamic solution concept for abstract games. *Journal of Optimization Theory and Applications* 32: 151–169.
Shubik, M. 1959. Edgeworth market games. In *Contributions to the theory of games IV, annals of mathematical studies 40*, ed. F.R. Luce and A. Tucker. Princeton: Princeton University Press.
Shubik, M. 1971. The bridge game economy: An example of indivisibilities. *Journal of Political Economy* 79: 909–912.
Shubik, M., and M. Wooders. 1982. Near markets and market games. Cowles Foundation Discussion Paper No. 657, also published as 'Clubs, near markets and market games'. In *Topics in mathematical economics and game theory: Essays in honor of Robert J. Aumann*, ed. M. Wooders. Fields Institute Communication 23. Providence: American Mathematical Society, 2000.
Shubik, M., and M. Wooders. 1983a. Approximate cores of replica games and economies: Part I. Replica games, externalities, and approximate cores. *Mathematical Social Sciences* 6: 27–48.
Shubik, M., and M. Wooders. 1983b. Approximate cores of replica games and economies: Part II. Set-up costs and firm formation in coalition production economies. *Mathematical Social Sciences* 6: 285–306.
Tauman, Y., A. Urbano, and J. Watanabe. 1997. A model of multiproduct price competition. *Journal of Economic Theory* 77: 377–401.
Tiebout, C. 1956. A pure theory of local expenditures. *Journal of Political Economy* 64: 416–424.
Tijs, S. 1981. Bounds for the core and the tau-value. In *Game theory and mathematical economics*, ed. O. Moeschlin and D. Pallaschke. Amsterdam: North-Holland.
van den Nouweland, A. 2005. Models of network formation in cooperative games. In *Group formation in economics: Networks, clubs and coalitions*, ed. G. Demange and M. Wooders. Cambridge: Cambridge University Press.
von Neumann, J., and O. Morgenstern. 1953. *Theory of games and economic behavior*. Princeton: Princeton University Press.
Vohra, R. 1999. Incomplete information, incentive compatibility and the core. *Journal of Economic Theory* 86: 123–147.
Volij, O. 2000. Communication, credible improvements and the core of an economy with asymmetric information. *International Journal of Game Theory* 29: 63–79.
Watts, A. 2001. A dynamic model of network formation. *Games and Economic Behavior* 34: 331–341.
Weber, S. 1979. On ε-cores of balanced games. *International Journal of Game Theory* 8: 241–250.
Weber, S. 1981. Some results on the weak core of a non-sidepayment game with infinitely many players. *Journal of Mathematical Economics* 8: 101–111.
Weber, S., and H. Wiesmeth. 1991. The equivalence of core and cost share equilibria in an economy with a public good. *Journal of Economic Theory* 54: 180–197.
Wilson, R. 1978. Information, efficiency and the core of an economy. *Econometrica* 46: 807–816.
Winter, E., and M. Wooders. 1994. An axiomatization of the core for finite and continuum games. *Social Choice and Welfare* 11: 165–175.
Wooders, M. 1978. Equilibria, the core, and jurisdiction structures in economies with a local public good. *Journal of Economic Theory* 18: 328–348.
Wooders, M. 1980. The Tiebout hypothesis: Near optimality in local public good economies. *Econometrica* 48: 1467–1486.
Wooders, M. 1983. The epsilon core of a large replica game. *Journal of Mathematical Economics* 11: 277–300.
Wooders, M. 1994a. Equivalence of games and markets. *Econometrica* 62: 1141–1160.
Wooders, M. 1994b. Large games and economies with effective small groups. In *Game-theoretic methods in general equilibrium analysis*, ed. J.-F. Mertens and S. Sorin. Dordrecht: Kluwer Academic Publishers.
Wooders, M., and W.R. Zame. 1987. Large games; fair and stable outcomes. *Journal of Economic Theory* 42: 59–93.
Xue, L. 1998. Coalitional stability under perfect foresight. *Economic Theory* 11: 603–627.
Yannelis, N. 1991. The core of an economy with differential information. *Economic Theory* 1: 183–198.
Yi, S.-S. 1997. Stable coalition structures with externalities. *Games and Economic Behavior* 20: 201–237.

Coase Theorem

Francesco Parisi

Abstract

The Coase Theorem holds that, regardless of the initial allocation of property rights and choice of remedial protection, the market will

determine ultimate allocations of legal entitlements, based on their relative value to different parties. Coase's assertion has occasioned intense debate. This article provides an intellectual history of Coase's fundamental theorem and surveys the legal and economic literature that has developed around it. It appraises the most notable attacks to the Coase Theorem, and examines its methodological implications and normative and practical significance in legal and policy settings.

Keywords

Adverse selection; American Law and Economics Association; Asymmetric information; Bargaining; Coase theorem; Contract enforcement; Efficient allocation; Entropy; Externalities; Free rider problem; Hold-up; Inalienability; Incentives; Income effect; Law, economic analysis of; Liability rules; Pigou, A. C.; Pigouvian taxes; Plant, A.; Private information; Property fragmentation; Property rights; Public goods; Scarcity; Social cost; Stigler, G.; Strategic behaviour; Tort; Transaction costs; Voluntary transfers

JEL Classifications

D62

Mutuality of advantage from voluntary exchange is one of the most fundamental concepts in economics. The well-known proposition of Ronald H. Coase (1960) – generally known as the Coase theorem – builds on this simple and yet fundamental insight. The law creates many rights and legal entitlements, establishing the initial allocation of rights and liabilities. Whenever there are no legal or factual impediments to exchange, the dynamic of the market will determine the final allocation of such rights.

In this context, Coase suggests that the transferability of rights in a free economy leads toward their best use and an efficient final allocation. Whenever the initial allocation is not optimal, the owners of the rights will have an incentive to transfer them to other individuals who value them more. Such an exchange will continue until there is no further potential for reciprocal profit, which will not be exhausted until each right is in the hands of the highest-valuing individual. The Coase theorem predicts that, in a competitive market environment without legal or factual impediments to exchange, the final allocation of rights will be efficient.

This article discusses the pervasive methodological implications of Ronald Coase's idea to the field of law.

A Brief Intellectual History

Coase's assertion that an initial assignment of property rights is often irrelevant to overall welfare has occasioned one of the most intense and fascinating debates in the history of legal and economic thought. Private property is often explained as the unavoidable by-product of scarcity in a world where common pool losses outweigh the sum of contracting costs and enforcement of exclusive property rights. At the turn of the 20th century, the underlying assumption in the economic literature was that private property emerged out of a spontaneous evolutionary process because of the desirable features of private property regimes in the creation of incentives for constrained optimization.

This understanding of the relationship between scarcity and emergence of legal entitlements characterized mainstream property right theory when Coase entered the academic world. Coase began his undergraduate studies at the London School of Economics in 1929, as a candidate for a Bachelor Degree in Commerce. In those years, one of Coase's teachers, Sir Arnold Plant, was re-examining the theme of property rights from a novel perspective. According to Plant, the traditional justification for private property – scarcity – was incapable of serving as the sole intellectual foundation for this institution. Plant showed that incentives, rather than scarcity, lay at the core of the property right problem (Plant 1974).

Coase's use of legal rules as an object of economic research in his analysis of incentive structure and alternative final resource allocations reveals a remarkable technical affinity with the

work of his undergraduate teacher. In his Nobel memorial lecture, Coase acknowledges the importance of his encounter with Plant as a 'great stroke of luck' that cultivated his interest in property rights theory (Coase 1992, p. 715). For Coase, Plant's teaching that '[t]he normal economic system works itself' (Salter 1921, pp. 16–17) and that prices in a competitive market lead resources to their highest valuing uses was a revelation into the dynamic of the economic system: 'I was then 21 years of age, and the sun never ceased to shine. I could never have imagined that these ideas would become some 60 years later a major justification for the award of a Nobel Prize. And it is a strange experience to be praised in my eighties for work *I* did in my twenties' (Coase 1992, p. 716).

The experience of the following years at the London School of Economics laid the methodological foundations of what would later become Coase's theorem on the problem of social costs. All the ingredients of his revolutionary analysis on the debated theme of social cost had been profiled during his LSE years (see Williamson and Winter 1991, pp. 34–5). But it is not until the late 1950s that Coase verbalized such a simple and yet ingenious idea. He had first expounded the core of his later theorem in an article published in 1959. In those pages, one grasps what would later become the central theme of Coase's celebrated argument:

> Whether a newly discovered cave belongs to the man who discovered it, the man on whose land the entrance to the cave is located, or the man who owns the surface under which the cave is situated is no doubt dependent on the law of property. But the law merely determines the person with whom it is necessary to make a contract to obtain the use of the cave. Whether the cave is used for storing bank records, as a natural gas reservoir, or for growing mushrooms depends, not on the law of property, but on whether the bank, the natural gas corporation, or the mushroom concern will pay the most in order to be able to use the cave. (1959, p. 25)

The discussion of the rationale of property rights under Coase's highest bidder framework obviously contained an attack on the Pigouvian approach (Pigou 1920) to the problem. The point was rather self-evident to Coase, but not so for some of the Chicago economists. George Stigler was among Coase's early critics:

> Ronald Coase criticized Pigou's theory rather casually, in the course of a masterly analysis of the regulatory philosophy underlying the Federal Communication Commission's [FCC] work. Chicago economists could not understand how so fine an economist as Coase would make so obvious a mistake. Since he persisted, we invited Coase (he was then at the University of Virginia) to come and give a talk on it. Some twenty economists from Chicago and Ronald Coase assembled one evening at the home of Aaron Director. . . . In the course of two hours of argument the vote went from twenty against and one for Coase to twenty-one for Coase. What an exhilarating event! (Stigler 1988, pp. 75–6)

According to Coase, the objections to his FCC paper are at the origin of his later 1960 article on the problem of social costs. Coase recalls that he was urged to omit that section of his FCC article, something he refused to do. In retrospect, Coase believes that had it not been for the Chicago economists' attacks his full-fledged idea would have never been formulated (1993, p. 250).

The Positive Coase Theorem

The arguments that were refined in the course of such debate were later put together in the form of an article for the *Journal of Law and Economics* in 1960, titled 'The Problem of Social Cost'. This article – later known as the Coase theorem – soon became a milestone in legal and economic literature. In the course of his austere discussion, Coase does not reveal any sign of anticipated realization of the revolutionary power of his insight. Indeed, Coase insists that he never intended to convey his thoughts in the precise and analytical form of a theorem (1988, p. 157).

A few years after the publication of 'The Problem of Social Cost', a sizeable number of commentaries and theoretical elaborations were developed on Coase's newly presented theme. The unpretentious style of Coase's article had thus been crowned by a notoriety rarely attained by legal writings of any sort (Shapiro 1985, p. 1540). Part of the uproar is explained by the fact that the article challenged an established

principle of public finance (see Manne 1975, pp. 123–6). Before 'The Problem of Social Cost', very little attention had been given to the possibility that the problem of externalities could be resolved through free market exchanges.

Coase boldly attacked the conclusions reached by the Pigouvian tradition by suggesting its influence was in part due to the lack of clarity in its exposition (1960, p. 39). Coase departs from the Pigouvian approach by demonstrating that, in the absence of transaction costs, generators and victims of externalities will negotiate an efficient allocation of resources, independent of the initial assignment of rights among them. In confuting the conclusions of the Pigouvian tradition, Coase gave life to a model with the potential for the evaluation of an unlimited number of legal and social issues.

George Stigler was the first scholar to restate Coase's model in the form of a theorem: '[U]nder perfect competition private and social costs will be equal' (1966, p. 113). Demsetz (1967, p. 349) defined the theorem in the following terms: 'There are two striking implications of this process that are true in a world of zero transaction costs. The output mix that results when the exchange of property rights is allowed is efficient and the mix is independent of who is assigned ownership (except that different wealth distributions may result in different demands)'. Soon thereafter, Guido Calabresi stated the same principle more descriptively: 'Thus, if one assumes rationality, no transaction costs, and no legal impediments to bargaining, all misallocations of resources would be fully cured in the market by bargains' (Calabresi 1968, p. 68).

The implicit premise of Coase's analysis draws upon a fundamental postulate of microeconomic theory: the free exchange of goods in the market moves goods towards their optimal allocation. The voluntary transfer of individual rights in the marketplace, thus, will cure a non-optimal allocation of legal entitlements.

The Coasean Methodological Revolution

Coase's article constitutes, according to many commentators, the first example of an economic analysis of law in North American literature. The novelty of his approach inspired an entire generation of scholars – pioneers in this new branch of applied economics. Only a few months prior to receiving the Nobel Prize for economics, in occasion of the First Annual Meeting of the American Law and Economics Association, Ronald H. Coase was recognized, together with Guido Calabresi, Henry G. Manne and Richard A. Posner, as a founding father of Law and Economics. This recognition follows many years of challenging debate. Many of the writings that developed around 'The Problem of Social Cost' tested the premises of Coase's model, seeking to undermine the conditions of his model and stressing the lack of practical reach of his analysis.

Further criticisms pertained to three fundamental points. One group of critics observed that the Coase Theorem disregarded the inter-industrial long-term effects of the system (Calabresi 1965; Wellisz 1964). These critics argued that Coase ignored the possible disequilibria which may occur after the negotiation and the likely dynamic changes in the initial equilibrium. In the context of Coase's well-known example, if the right has been assigned to the ranchers, the farmer will have to pay local ranchers until they all relinquish their right of pasture. The entire cost will, thus, burden the farming industry. Farmers will either have to bear the burden of the injury caused by the livestock or agree to pay the price demanded by the ranchers, whichever is less, on the assumption that negotiation is costless. Under this liability rule, the cost of ranching will not reflect the cost imposed on the farmers. The transfer of rights and liability from one group to another will, therefore, result in a shift in the relative wealth and costs associated with the two industries. The criticism claims that, in the long run, every shift of wealth will lead to an inter-industrial disequilibrium.

In 1968, Calabresi, one of the initial proponents of this criticism, reconsidered it, noting that in the presence of determined conditions the conclusions of Coase remain as true in the long run as in the short term (1968, p. 67). Calabresi's later analysis re-established the authority of the Coase Theorem, at least on this point. It became clear that Coase did not ignore the long-term

effects of his model. Perhaps not explicitly, he had considered them to their logical extreme. Calabresi observes: 'The reason is simply that (on the given assumptions) the same type of transactions which cured the short run misallocation would also occur to cure the long run ones. . . . This process would continue until no bargain could improve the allocation of resources' (1967, pp. 67–8).

In 1972, Harold Demsetz joined this debate, demonstrating with a more systematic analysis that the conclusions reached by Coase are not corroded by the long-term effects of a change in the assignment of property rights. Demsetz's reasoning finds its basis in the principle that the process of allocation of scarce resources among alternative uses is analogous to the process of constrained optimization of the single owner of two conflicting activities.

An additional critique, formulated by Calabresi (1965) and Wellisz (1964), suggests that strategic behaviour in the bargaining process risks compromising Coase's results. These authors observe that the change in the rule of law creates the conditions for possible extortion on the part of the right holders against the other individuals who are bound by the rule. The argument is that individuals are likely to threaten the use of their own rights in a measure which exceeds the optimal level, in order to maximize the gain from the release of their own legal entitlements. By introducing the possibility of strategic behaviour in the negotiation, the result may differ from the optimal equilibrium. Demsetz (1972, p. 21) supplied a convincing answer to this criticism. According to Demsetz, the possibility of strategic behaviour in the negotiations does not alter the efficiency in the final allocation of resources between the two activities. Strategies will be capable of altering the internal distribution of the contractual surplus between the parties, but not the final outcome of the negotiation.

It should be noted, however, that the entire analysis presupposes that the so-called income effect can be ignored. In general, a different allocation of property rights implies a different distribution of wealth between the individuals involved. Different initial endowments generate different final allocations, notwithstanding an equal level of efficiency. In order for the final allocations to be identical, it is necessary that the utility functions of the individuals involved are almost linear. The absence of the income effects implies, in this sense, that the demand functions for the good are independent of the income level.

It should be further observed that the credibility of the threat made in the course of strategic bargaining finds its limits in the market structure in which the Coasean negotiation takes place. In general, the competitive structure of the market eliminates much of the advantage that can be obtained through strategic behavior in the negotiation process. Inasmuch as the market of resources is competitive, strategic bargaining is not capable of bringing about any abnormal return.

The criticism, however, appears to be on the mark when it argues that, in some marginal situations, the curing role of the free exchange may still be impeded. For example, consider reversing the assignment of property rights between the rancher and the farmer. In such a situation, the farmer is likely not to have an equally large number of alternatives. The transfer of a farm from one place to another is costly, and farming unavoidably requires the undertaking of location-specific investments. Since some capital investment is irreversibly locked in that specific location, the farmer has less opportunity to relocate than the rancher. The rancher, consequently, finds himself in a position of local monopoly in the sale of his property right. Demsetz considers the monopoly that affects this feature of the Coasean exchange identical to the standard monopoly of microeconomic analysis (1972, p. 24). According to Demsetz, the concerns for possible monopolistic structures in the market of rights considered by Coase must not, however, be used to raise again the already resolved problem of the initial allocation of rights, since reversing the rule of liability would simply result in the farmer now having monopoly power (1972, pp. 24–5).

A second group of critics concentrated on the distributive effects of the model (Regan 1972; Nutter 1968). They argued that a final efficient allocation of resources requires transfers of wealth

induced by the changed legal rule. Further, these critics observed that, even if one disregards the distributive effects of the rule, a different assignment of the right could in some cases create the conditions for strategic behaviour in negotiation capable of disturbing the efficiency of the final allocation.

A third group of authors focused on the scarce realism of the no-transaction-cost assumption (see Cooter 1987, p. 457). According to this criticism, the true Achilles' heel of Coase's analysis was in the unrealistic assumption of absence of costs in the process of negotiation and transfer of the right. These authors observed that the idea of a transaction without cost is a logical fiction cloaking a mere tautology.

The Normative Coase Theorem

The utility of models predicting behaviour in a zero transaction-cost world is that they guide the law – whose object is to develop rules which approximate the zero transaction-cost world as closely as possible – in responding to legal problems arising in a positive transaction-cost environment (Epstein 1993). The vast literature that developed around Coase's theorem formulated important normative corollaries of it, based on the evaluation of the relative costs of alternative assignments of rights.

According to the positive Coase theorem, absent transaction costs, the final allocation of scarce resources would coincide with the use that an individual who is the single owner of different activities would make of his endowments, regardless of the initial assignment of rights and choice of remedial protection. When transaction costs are present, however, an exchange will be pursued only to the point at which its marginal benefit equals the marginal cost of the transaction. If transaction costs exceed the benefits of a contract, no exchange will take place in the market. For a right to be exchanged it is necessary that transaction costs be less than the difference between the demand and supply prices. If this condition is not met, the Coasean bargaining will not be carried out, and both initial assignment of rights and choice of remedies will affect final allocations.

The Relevance of Transaction Costs and the Simple Normative Coase Theorem

The notion of transaction costs has acquired particular importance in law and economics as the absence of transaction costs represents a fundamental condition for the applicability of the positive Coase theorem. Although at first impression transaction costs play a role analogous to transportation costs in international trade or, more generally, to the contracting costs in the economics of exchange (Demsetz 1972, p. 20), in Coase's world the role of transaction costs has much greater normative implications.

For purposes of the theorem, the notion of transactions costs should include not only bargaining costs associated with the negotiation and conclusion of the contract but also all costs associated with the strategic behaviour of the parties and the execution and enforcement of the transaction. The notion of transaction costs should thus include *ex ante* costs due to asymmetric information, adverse selection, free riding, and hold-up strategies, as well as *ex post* costs associated with monitoring and enforcing the contracts.

Strategic behaviour may be an important source of transaction costs in a Coasean setting. In Coase's various examples, the property rights which are exchanged are private goods, characterized by their excludability. Difficulties arise when the object of the Coasean bargaining is an entitlement which has the nature of a public good (see Cheung 1970, pp. 49–70). Due to the well-known problems associated with the supply of public goods, the Coasean bargaining solution may fail to cure a non-optimal allocation of rights that falls within this category. Consider a scenario in which the object of the Coasean negotiation consists of a non-excludable right (for example, the right to enjoy pollution-free air in a residential environment). As well known, individuals will not reveal their own preferences for public goods through the price system, placing public goods among those cases that are most resistant to the Coasean antidote.

A first simple normative reformulation of the Coase theorem focuses on transaction costs and the role that legal systems may play in reducing these impediments to voluntary bargaining. Legal rules can lower obstacles to private bargaining, such as by reducing transaction costs and minimizing other costs associated with transfer (strategic, legal, and so on). For this reason, transactional cost considerations should be fundamental to any analysis of legal regimes and the design of contracting processes, governance mechanisms and institutions.

The Complex Normative Coase Theorem

The first original formulation of Coase's proposition can be restated as a normative theorem: in the presence of positive transaction costs, the efficiency of the final allocation is not independent of the choice of the legal rule, and that the preferable initial assignment of rights is that which minimizes the effects of such transaction costs. The various normative restatements of the Coase Theorem aim at identifying legal rules and remedies that replicate the outcomes of a hypothetical Coasean bargaining or to mimic the solution that would be chosen by the single owner of interfering resources.

Important normative reformulations of the Coase Theorem focus on two important elements: relevance of initial assignment of rights and relevance of remedial protection. Demsetz (1972) and Calabresi and Melamed (1972) were among the first to discuss systematically the problems resulting from lifting the assumption of zero transaction costs. Articulating the normative core of the Coase theorem, Demsetz observes that the introduction of significant transaction costs into the choice of liability rule analysis does affect resource allocation. One liability rule may be superior to another because the difficulty of avoiding costly interactions is usually different for the interacting parties. Accordingly, the normative predicament indicates that the rule of liability should be based on which party can avoid the costly interaction at the lowest cost.

When two or more parties have conflicting interests in the same resource, the law must decide which party shall prevail, that is, which party shall receive the entitlement. Once the entitlement decision is made, the law must decide how the entitlement is to be protected and whether it may be transferred. Articulating a concept of entitlements protected by property, liability or inalienability rules, Calabresi and Melamed (1972) develop a framework that integrates the approaches of property and tort. Entitlements can be protected by property rules (transfer of the entitlement involves a voluntary sale by its holder), liability rules (the entitlement may be destroyed by another party if he is willing to pay an objectively determined value for it), or rules of inalienability (transfer of the entitlement is not permitted, even between a willing seller and a willing buyer). Calabresi and Melamed allow for a wide range of concerns to be balanced through the assignment of a particular entitlement. Calabresi and Melamed outline how, given the reality of transaction costs, an economic efficiency approach selects one allocation of entitlements over another. Entitlements cannot be enforced solely through property rules because, even if the transfer would benefit all parties, high transaction costs (especially the hold-up problem) may prevent an efficient reallocation. Calabresi and Melamed demonstrate how liability rules often achieve a combination of efficiency and distributive results that would be difficult to achieve under a property rule. Calabresi points out that Coase's analysis offers invaluable instruments for the identification of the areas in which public intervention becomes desirable (Calabresi 1968, pp. 72–3). In its normative version, the theorem indicates that legal rules that minimize the effects of such costs are to be preferred for being relatively more efficient (Polinksy 1989, p. 14). In its more complex formulation, the Coase theorem provides, indeed, a guide for such a choice.

The following is a classic illustration (Polinsky 1989, pp. 11–14). The smoke of a factory soils laundry which is line drying on five neighbouring properties. The losses amount to $150 for each neighbour, for a total of $750. The damage could be eliminated through the installation of a purifying filter on the industrial smokestack or through the acquisition of electric dryers on the part of each one of the neighbouring owners. The cost

of the filter would amount to $300, while the dryers would impose a cost of $100 per household, for a total of $500. The first solution is obviously more efficient, since the acquisition of five dryers would require a greater expenditure than the single filter. The Coase theorem predicts that in the absence of transaction costs the efficient solution will be chosen independently of the initial assignment of property rights. Even if we assume an initial allocation of polluting right to the industry (that is, fully legalizing industrial emissions), the landowners would jointly offer to buy the industrial filter at their expense. Sharing the cost of the filter in equal parts, each owner would face a cost of only $60, with a relative saving of $40 compared with the otherwise necessary acquisition of a personal dryer.

If we relax the initial assumption of no transaction costs, the initial allocation of property rights no longer is immaterial. Imagine that each owner has to face a cost of $120 in order to negotiate the contract with his neighbours and with the owner of the industrial plant. If the right is assigned to the industry, each landowner will have to choose whether to bear the loss of his soiled laundry for $150, to acquire the electric dryer for $100, or, finally, to undertake the negotiation process for a total pro-rata cost of $180. Considering these alternatives, each rational landowner would choose to acquire his own dryer, generating a socially non-optimal outcome. However, the assignment of property rights to the neighbouring residents rather than to the polluting industry would minimize the effect of positive transaction costs, since the industry would have incentives to install the filter, without any need for Coasean bargaining with the neighbours.

Two impediments to bargaining (that is, sources of transaction costs) take the form of externalities and hold-up, which Epstein (1993) shows stand in inverse relationship to each other. He defines the optimal legal rule as that which minimizes the sum of these externality and holdout costs in any particular institutional setting. Epstein demonstrates, through examples in property, restitution and tort, how Coase's transaction costs model plays the central organizing role in developing legal responses to many private law problems. Notwithstanding the obvious measurement and information problems, Epstein (1993) stresses the importance of the 'single owner test': where resources are under the command of two or more persons, the legal arrangement should attempt to induce all the parties to behave in the same way that a single owner would. Epstein concludes that, where the single owner test yields a unique result, that result should be adopted as the legal rule. Where the single owner test does not yield clear results, however, no corollary principle will provide a decisive answer to the particular problem.

Further exploring the choice between property and liability rules suggested by Calabresi and Melamed, Kaplow and Shavell (1996) address several factors casting doubt on the equivalence of these alternatives in low transaction-cost environments. Their analysis considers several objections to Coasean costless bargaining, including the inability of a party to ascertain what the other is willing to pay or accept, victims' ability to mitigate harm, the problem posed by one party being judgment proof, and administrative costs. Kaplow and Shavell find a presumption in favour of liability rules over property rules in the context of harmful externalities, but that this may be overcome as a result of one of more of the factors they describe. After considering some of the proffered justifications for the use of property rules to protect possessory interests, the authors find a strong theoretical case for the protection of these interests using property rules. The normative Coase theorem thus underlies the choice of the optimal system to ensure the protection of various types of property rights.

Also bridging the gap between Coase, where liability rules and property rules are equally efficient, and Calabresi and Melamed, where high transaction costs lead to a preference for liability rules, is the work by Ayres and Talley (1995) on private information as a transaction cost. The inefficiency occurs when parties misrepresent their own valuations to gain strategic advantage in the bargaining process. Focusing on the effect of splitting an entitlement between two rivalrous users rather than among buyers or among sellers, these authors find that, when two parties have

private information about how much they value an entitlement, endowing each party with a partial claim to the entitlement can reduce the incentive to behave strategically during bargaining by inducing greater disclosure. A bargainer has two Coasean alternatives: buy the other party's claim or sell one's own claim. The normative formulation of Ayres and Talley is that a liability rule regime is preferable because it allows a party's decision to pursue one of these alternative transactions to function as a credible signal of a low or high valuation, thereby encouraging more efficient trade.

Building upon the literature on property fragmentation (Heller 1998; Buchanan and Yoon 2000), Parisi (2002) and Schulz et al. (2002) suggest that property is subject to a fundamental law of entropy. In the property context, entropy induces a one-directional bias. This bias is driven by asymmetric transaction costs – it is often harder to reunite separated property bundles than to break them apart. Parisi hypothesizes that courts and legislators account for the presence of asymmetric transaction costs and correct for problem through the selective use of remedies and by selecting default rules designed to minimize the total deadweight losses of property fragmentation. Parisi (2006) offers a reformulation of the normative Coase theorem in situations characterized by asymmetric transaction and strategic costs, such as when complementary fragments of property are attributed to different owners. The asymmetry arises from the fact that it is often harder to reunite separated property bundles than to break them apart. This variant of the Coase theorem turns on (*a*) an initial allocation of entitlements that minimizes the effects of the positive transaction costs, and (*b*) the selection of legal rules that reduce social welfare losses by facilitating optimal levels of reunification.

The Coase Theorem and Its Legacy in Law and Economics

In 1960 Coase entrusted legal and economic scholars with the challenging task of deriving the implications of his theorem in their areas of research. Coase's invitation was taken up by a number of economists and lawyers who experimented with the unparalleled analytical potential of Coase's theorem in their research. According to Coase, economists in the Pigouvian tradition fail to consider the possible reciprocity of the effects of individual choices. By labelling one agent as injurer and the other as victim, the Pigouvian tradition presumes an initial allocation of rights (Cornes and Sandler 1986, p. 59). In such a manner this approach falls into a serious methodological error, notwithstanding empirical psychological studies suggesting otherwise (see Kahneman et al. 1990, pp. 1325–48). By taxing the generator of the externality in a measure corresponding to the difference between the private cost and the social cost of his own activity, the followers of Pigou fail to consider the effects of potential victims' behaviour. If the social cost of the industrial emissions is calculated by aggregating the economic disadvantages of the residents who are negatively affected by the smoke, the figure will vary with the number of individuals who fix their residence in that area. If the Pigouvian tax is imposed on the industrial activity only, there will be less incentive for each resident to consider moving into a different neighbourhood. New individuals may actually locate their residence in that area, without considering the potential increase in the costs imposed on the industrial activity.

Through these arguments, Coase's analysis demonstrates the incapacity of the Pigouvian approach to consider the interdependence of the harmful effects generated by individual choices. Coase's analysis occasioned a paradigmatic shift in legal and economic analysis, and, as Henry Manne once observed, 'it is hard to imagine law ever again being free of the influence of the techniques and findings of objective economic analysis' (1993, p. 4). His theorem, short of providing a simplistic formula for the social cost problem, suggests an alternative approach based on the evaluation of the relative costs of alternative assignments of rights and legal protection.

See Also

- ▶ Hold-Up Problem
- ▶ Property Rights

Bibliography

Ayres, I., and E. Talley. 1995. Solomonic bargaining: Dividing a legal entitlement to facilitate Coasean trade. *The Yale Law Journal* 104: 1027–1117.

Buchanan, J., and Y.J. Yoon. 2000. Symmetric tragedies: Commons and anticommons property. *Journal of Law and Economics* 29: 1–13.

Calabresi, G. 1965. The decision for accidents: An approach to non-fault allocation of costs. *Harvard Law Review* 78: 713–745.

Calabresi, G. 1968. Transaction costs, resource allocation and liability rules: A comment. *Journal of Law and Economics* 11: 67–68.

Calabresi, G., and A.D. Melamed. 1972. Property rules, liability rules, and inalienability: One view of the cathedral. *Harvard Law Review* 85: 1089–1128.

Cheung, S. 1970. The structure of a contract and the theory of a non-exclusive resource. *Journal of Law and Economics* 13: 49–70.

Coase, R.H. 1959. The Federal Communications Commission. *Journal of Law and Economics* 2: 1–40.

Coase, R.H. 1960. The problem of social cost. *Journal of Law and Economics* 3: 1–44.

Coase, R.H. 1988. *The firm, the market, and the law*. Chicago: University of Chicago Press.

Coase, R.H. 1992. The institutional structure of production. *American Economic Review* 82: 713–719.

Coase, R.H. 1993. Law and economics at Chicago. *Journal of Law and Economics* 36: 239–254.

Cooter, R. 1987. Coase theorem. In *The new Palgrave: A dictionary of economics*, ed. J. Eatwell, M. Milgate, and P. Newman, vol. 1. London: Macmillan.

Cornes, R., and T. Sandler. 1986. *The theory of externalities, public goods, and club goods*. Cambridge: Cambridge University Press.

Demsetz, H.M. 1967. Toward a theory of property rights. *American Economic Review* 57: 347–359.

Demsetz, H.M. 1972. When does the rule of liability matter? *The Journal of Legal Studies* 1: 13–28.

Epstein, R.A. 1993. Holdouts, externalities, and the single owner: One more salute to Ronald Coase. *Journal of Law and Economics* 36: 553–586.

Heller, M.A. 1998. The tragedy of the anticommons: Property in the transition from Marx to markets. *Harvard Law Review* 111: 621–688.

Kahneman, D., J.L. Knetsch, and R.H. Thaler. 1990. Experimental tests of the endowment effect and the Coase theorem. *Journal of Political Economy* 98: 1325–1348.

Kaplow, L., and S. Shavell. 1996. Property rules versus liability rules: An economic analysis. *Harvard Law Review* 109: 723–754.

Kuhn, T.S. 1970. *The structure of scientific revolutions*. 2nd ed. Chicago: University of Chicago Press.

Manne, H.G. 1975. *The economics of legal relationships*. St. Paul: West Publishing Co.

Manne, H.G. 1993. *The intellectual history of George Mason University School of Law*. Arlington: School of Law, George Mason University.

Nutter, G.W. 1968. The Coase theorem on social cost: A footnote. *Journal of Law and Economics* 16: 503–507.

Parisi, F. 1995. Private property and social costs. *European Journal of Law and Economics* 2: 149–173.

Parisi, F. 2002. Entropy in property. *American Journal of Comparative Law* 50: 595–632.

Parisi, F. 2006. Entropy and the asymmetric Coase theorem. In *Property rights dynamics: A law and economics perspective*, ed. D. Porrini and G. Ramello. London: Routledge.

Pigou, A.C. 1920. *The economics of welfare*. London: Macmillan.

Plant, A. 1974. *Selected economic essays and addresses*. London: Routledge and Kegan Paul.

Polinsky, A.M. 1989. *An introduction to law and economics*. Boston: Little, Brown and Company.

Regan, D.H. 1972. The problem of social cost revisited. *Journal of Law and Economics* 15: 427–433.

Salter, J.A. 1921. *Allied shipping control*. Oxford: Clarendon Press.

Schulz, N., F. Parisi, and B. Depoorter. 2002. Fragmentation in property: Towards a general model. *Journal of Institutional and Theoretical Economics* 158: 594613.

Shapiro, F.R. 1985. The most cited law review articles. *California Law Review* 73: 1540–1554.

Stigler, G.J. 1966. *The theory of price*. 3rd ed. New York: Macmillan.

Stigler, G.J. 1988. *Memoirs of an unregulated economist*. New York: Basic Books.

Wellisz, S. 1964. On external diseconomies and the government assisted invisible hand. *Economica* 31: 345–362.

Williamson, W., and S.G. Winter. 1991. *The nature of the firm: Origins, evolution, and development*. New York: Oxford University Press.

Coase, Ronald Harry (Born 1910)

Steven G. Medema

Abstract

Ronald Coase made seminal contributions to law and economics and to the theory of the firm, for which he received the 1991 Nobel Prize. The importance of understanding the

role of transaction costs in economic activity and the influence of alternative institutional structures on economic performance are hallmarks of Coase's scholarship, and both the economic analysis of law and the new institutional economics are outgrowths of his work. Coase occupies a significant although somewhat controversial place in the history of the Chicago School of economics.

Keywords

Accounting; Average cost pricing; Bargaining; Chicago School; Coase conjecture; Coase theorem; Coase, R. H.; Cobweb theorem; Consumer theory; Externalities; Firm, organization of; Government failure; Imperfect competition; Knight, F.; Law and economics; Lerner, A.; Marginal cost pricing; Market failure; Mathematics and economics; Monopoly; Multi-part pricing; New institutional economics; Opportunity cost; Posner, R.; Public goods; Public utilities; Public utility pricing; Rational expectations; Social cost theory; Special interests; Transaction costs; Viner, J.

JEL Classifications

B31

Ronald Harry Coase was born on 29 December 1910 in the London suburb of Willesden. He received the BSc in Commerce from the London School of Economics in 1932 and while there was greatly influenced by Arnold Plant, who, as Coase has said, taught him many of the lessons that later came to be associated with the Chicago School. Interestingly, Coase did not take a single economics course while he was at the LSE, which he suggests gave him 'a freedom in thinking about economic problems which [he] might not otherwise have had' (1990, p. 3).

Upon completing his studies at the LSE, Coase took up a position at the Dundee School of Economics and Commerce, where he taught with his friend and public choice pioneer Duncan Black from 1932 to 34. Coase moved on to the University of Liverpool in 1934–35 before returning to the LSE, where he remained from 1935 until 1951. His time at the LSE was interrupted by the Second World War, during which he served as a statistician at the Forestry Commission (1940–41) and in the Central Statistical Office, Offices of the War Cabinet (1941–46). Coase left the LSE for the US and the University of Buffalo in 1951, remaining there until 1958. Following a year spent at the Center for Advanced Study in the Behavioral Sciences at Stanford, he accepted an appointment at the University of Virginia in 1959.

Although Coase is most closely associated with the Chicago School, his two most influential works – 'The Nature of the Firm' (1937) and 'The Problem of Social Cost' (1960) – were written before he arrived at Chicago, in 1964, to teach at the Law School and to join Aaron Director in editing the *Journal of Law and Economics.* Coase retired from the University of Chicago in 1981 and was awarded the Nobel Prize in Economics in 1991.

Scholarly Work

While most economists identify Coase with his two classic articles on the firm and social costs, his published output is very extensive and ranges across topics such as accounting, advertising, public goods, consumer surplus, public utility pricing, monopoly theory, blackmail, the economic role of government, and the history of economic thought. Several themes appear throughout Coase's work: the important role played by institutions – in particular the firm, the market and the law – in determining economic structure and performance, the role of transaction costs in economic activity, the need for a comparative institutional approach to economic policy, and a distaste for abstract theorizing. These themes come through unmistakably in *The Firm, the Market and the Law* (1988) and *Essays on Economics and Economists* (1994), which, together, collect many of Coase's most significant writings.

The lion's share of Coase's work during the first part of his career dealt, in one way or another, with firm behaviour and organization. His earliest contributions analysed the formation of

producers' expectations (for example, Coase and Fowler 1935), using the pig cycle as the case study. The conventional cobweb theorem explanation for these cycles was that producers expected current prices and costs to continue into the future. The adjustments in supply that resulted then gave rise to disequilibrium cycles. Coase and Fowler found that this explanation was incorrect – that producers did in fact adjust their expectations of prices and costs very quickly, and that the prediction errors arose from the difficulty of predicting variations in demand and in foreign supply. This work was later cited by J.F. Muth (1961, p. 21) in one of his classic papers on rational expectations. Coase also collaborated with Fowler and Ronald Edwards on a series of pieces dealing with the interrelations between accounting and economics (for example, Coase 1938; Coase et al. 1938). These writings, which were very much in the LSE cost tradition, demonstrated that traditional accounting practices do not adequately capture the true (opportunity) nature of costs and also pointed to the problematic nature of designing workable accounting methods to do so.

Coase also wrote a number of articles dealing with monopoly and imperfect competition, a few of which bear mention of here. Two of his theoretical pieces are of particular import. 'Durability and Monopoly' (1972) demonstrated that a monopoly firm which produces a good that is infinitely durable will be forced to sell the good at the competitive price, unless it can decrease the durability of the good or make contractual arrangements through which it promises to limit its production – a result which has come to be known as 'the Coase conjecture'. 'The Marginal Cost Controversy' (1946) is Coase's most significant work on monopoly and deals with public utility pricing and regulation. Abba Lerner and others had claimed that marginal cost pricing accompanied by a government subsidy is the efficient pricing policy for public utilities. Against this, Coase argued that marginal cost pricing is inferior to a system of multi-part pricing and may in fact be inferior to *average* cost pricing. This paper, and three related papers that followed it, are illustrative of one of the central themes in Coase's work – that, in assessing the efficiency of economic outcomes, one must focus broadly, rather than narrowly, on benefits, costs, and incentives.

Coase's work on public utilities also has an historical strand. Articles on the British Post Office discuss the rise of the penny postage in Great Britain under Rowland Hill and the attempts by the Post Office to enforce its monopoly against incursions by private entrepreneurs, including the messenger companies (for example, 1955). His study of British broadcasting analyses the development of wireless and wire radio broadcasting, as well as of television broadcasting and the rise of the BBC as the monopoly supplier of all of the above (1950, 1954). His interest in the government's role in broadcasting carried over to the United States and an analysis of the role of the Federal Communications Commission (1959, 1966) in the allocation of broadcast frequencies. In fact, it was from this study that 'The Problem of Social Cost' came to be written.

While the foregoing gives a sense for the breadth of Coase's contributions, it is unquestionable that his most influential work is contained in two papers – 'The Nature of the Firm' (1937) and 'The Problem of Social Cost' (1960), the two works cited by the Royal Swedish Academy in awarding Coase the Nobel Prize. In the former, Coase set out to explain why firms exist and what determines the extent of a firm's activities. He found the answer in a concept to which most economists had until recently paid scant attention – transaction costs. Coase suggested that we tend to see firms emerge when the cost of internal organization is lower than the cost of transacting in the market, and that the limit of a firm's activities (or, the extent of internal organization) comes at the point where the cost of organizing another transaction internally exceeds the cost of transacting through the market. Although published in 1937, 'The Nature of the Firm' attracted little attention until the early 1970s, when Oliver Williamson, Armen Alchian, Harold Demsetz and others began to build on or take off from Coase's contribution to bring transaction costs, the contracting process, and firm organization to the fore in economic analysis.

'The Problem of Social Cost' took the transaction-cost paradigm in a different direction –

the legal-economic arena and situations of conflicts over rights. Although 'The Problem of Social Cost' is one of the most cited articles in all of the economics and legal literatures, it has also been widely misunderstood. From this paper comes the now-famous Coase theorem – actually codified as such by George Stigler (1966) – which says that when transaction costs are zero and rights are fully specified, parties to a dispute will bargain to an efficient outcome, regardless of the initial assignment of rights. But Coase recognized that the transaction costs are pervasive and will generally preclude the working of this bargaining mechanism. Coase thus concludes that legal decision-makers should assign rights so as to maximize the value of output in society – a concept that lies at the heart of the modern law and economics movement (Medema 1999; Medema and Zerbe 2000).

The crux of 'The Problem of Social Cost,' however, is Coase's attempt to demolish the Pigovian tradition of social cost theory (Pigou 1932). The analysis that came to be known as the Coase theorem was used to demonstrate that, under standard neoclassical assumptions, Pigovian remedies for externalities are unnecessary: costlessly functioning markets, like the costlessly functioning governments of Pigovian welfare theory, will generate efficient outcomes. The problem, as Coase pointed out, is that neither markets nor governments function costlessly, and thus neither will generate optimal solutions. This leaves policymakers with a choice among imperfect alternatives, and Coase advocates a close examination of the benefits and costs associated with the alternative policy options, in order to facilitate the adoption of policies (including doing nothing at all) which maximize the value of output.

That government failure is at least as pervasive as market failure, and that economists are too quick to advocate tax, subsidy, and regulatory solutions without a careful examination of the situation, are recurring themes in Coase's work. His analyses of social cost issues, public utility pricing, and his classic article on role of the lighthouse in public goods theory as against the actual history of private lighthouse provision in Great Britain (1974) are excellent examples of Coase's position here. When Coase looks at government, he sees agencies captured by special interests, making policies that usually make matters worse rather than better, and operating in virtual ignorance of the virtues of the market. Yet a careful reading of Coase suggests that he is not 'anti-government' but, rather, an advocate for economic theorizing and policymaking which recognizes that policy choices are always among imperfect alternatives.

These criticisms are part of Coase's more general concern about the way that economists practice their trade (1994). He is suspicious of consumer theory as a whole and of the way in which mathematical and quantitative techniques have been used in modern economics. His own writings evidence some graphs and some technical intuitive analysis, but, reflecting Coase's lifelong distaste for using mathematics in his work, there is not an equation to be found. Coase believes that economists are obsessed with what he calls ' blackboard economics', an economics where curves are shifted and equations are manipulated on the blackboard, with little attention to the correspondence (or lack thereof) between these models and the real-world economic system. This, he says, has manifested itself in economists' ignorance of the role played by transaction costs and economic institutions generally, and in an approach to public policy that fails to examine in any kind of depth the consequences of alternative policy actions.

Coase and Chicago

Coase's critical attitude toward the practice of economics does not stop at the doors of the University of Chicago. Indeed, his close association with the Chicago School belies a degree of tension in the relationship and highlights the risks involved in thinking in terms of a homogeneous Chicago school. In spite of his position as a founding father of law and economics and, by extension, the expansion of the boundaries of economics so closely associated with Chicago, Coase has been critical of economic imperialism

generally and of the economic analysis of law in particular (Coase 1977, 1993). Coase's interest is not the economic analysis of law, but rather the study of how the legal system impacts the economic system – old-style Chicago law and economics of the sort being published in the *Journal of Law and Economics* in the 1960s and 1970s. As such, his interest and intellectual commonalities lie much more with the older Chicago school of Frank Knight and Jacob Viner than with the Becker–Stigler–Posner generation, and he has a much greater interest in the new institutional economics (of which he is also regarded as a founding father) than in the modern economic analysis of law movement à la Richard Posner. Coase has been chastised by Posner (1993) on this and other counts, but he remains unapologetic. That Coase has a place within the Chicago tradition goes without saying, but he has also remained his own man – dissenting from the received doctrine when it did not fit with his views.

See Also

- ▶ Chicago School
- ▶ Chicago School (New Perspectives)
- ▶ Coase Theorem
- ▶ Law, Economic Analysis of
- ▶ New Institutional Economics

Selected Works

1935. (With R.F. Fowler.) Bacon production and the pig-cycle in Great Britain. *Economica*, New Series 2: 142–167.

1937. The nature of the firm. *Economica*, New Series 4: 386–405.

1938. Business organisation and the accountant. In *Studies in costing*, ed. D. Solomons. London: Sweet and Maxwell, 1952.

1938. (With R.S. Edwards and R.F. Fowler.) *Published balance sheets as an aid to economic investigation*: *Some difficulties*. London: Accounting Research Association Publication No. 3.

1946. The marginal cost controversy. *Economica*, New Series 13: 169–182.

1950. *British broadcasting: A study in monopoly.* London: Longmans, Green and Co.

1954. The development of the British television service. *Land Economics* 30: 207–222.

1955. The postal monopoly in Great Britain: An historical survey. In *Economic essays in commemoration of the Dundee School of Economics 1931–1955,* ed. J.K. Eastham. London: William Culcross and Sons.

1959. The Federal Communications Commission. *Journal of Law and Economics* 2: 1–40.

1960. The problem of social cost. *Journal of Law and Economics* 3: 1–44.

1966. The economics of broadcasting and government policy. *American Economic Review* 56: 440–447.

1972. Durability and monopoly. *Journal of Law and Economics* 15: 143–149.

1974. The lighthouse in economics. *Journal of Law and Economics* 17: 357–376.

1977. Economics and contiguous disciplines. In *The organization and retrieval of economic knowledge,* ed. M. Perlman. Boulder: Westview Press.

1988. *The firm, the market, and the law.* Chicago: University of Chicago Press.

1990. Accounting and the theory of the firm. *Journal of Accounting and Economics* 12: 3–13.

1992. The institutional structure of production. *American Economic Review* 82: 713–719.

1993. Law and economics at Chicago. *Journal of Law and Economics* 36: 239–254.

1994. *Essays on economics and economists.* Chicago: University of Chicago Press.

Bibliography

Medema, S.G. 1994. *Ronald H. Coase*. London/New York: Macmillan/St. Martin's Press.

Medema, S.G. 1999. Legal fiction: The place of the Coase theorem in law and economics. *Economics and Philosophy* 15: 209–233.

Medema, S.G., and R.O. Zerbe Jr. 2000. The Coase theorem. In *The encyclopedia of law and economics*, ed. B. Bouckaert and G. De Geest. Aldershot: Edward Elgar Publishing.

Muth, J.F. 1961. Rational expectations and the theory of price movements. *Econometrica* 29: 315–335.
Pigou, A.C. 1932. *The economics of welfare*. 4th ed. London: Macmillan.
Posner, R.A. 1993. Ronald Coase and methodology. *Journal of Economic Perspectives* 7 (4): 195–210.
Stigler, G.J. 1966. *The theory of price*. 3rd ed. New York: Macmillan.
Williamson, O.E. 1975. *Markets and hierarchies: Analysis and antitrust implications*. New York: Free Press.

Cobb–Douglas Functions

Murray Brown

Abstract

Perhaps the most common form of production function in economics, the Cobb–Douglas function has a range of attractive properties. The input demand and supply of output functions have the property of continuous differentiability everywhere on their respective domains; and the form has a function coefficient that is identical to its degree of homogeneity, calculated by summing the factor production elasticities. Its restrictions have made it an object of disdain for some. But the Cobb–Douglas form is remarkably robust in a vast variety of applications and is therefore very likely to endure.

Keywords

Aggregation (production); CES production function; Cobb, C.; Cobb–Douglas functions; Douglas, P. H.; Elasticity of substitution; Factor substitution; Frontier production functions; Production functions; Technical change; Walras, L.; Wicksell, J. G. K.; Wicksteed, P. H.

JEL Classifications

E23

The Cobb–Douglas function is perhaps the most ubiquitous form in economics, owing its popularity to the exceptional ease with which it can be manipulated and to the fact that it possesses the minimal properties that economists consider desirable. It appeared early (at least by 1916; see Wicksell 1958, p. 133), notably in the theory of distribution where it was used to prove the adding-up theorem of factor shares when the production elasticities sum to unity. It is the first form that many embryonic mathematical economists squeeze and buffet to obtain nice expressions for marginal products and utilities. It has been applied econometrically countless times, still surprising people that it can explain the data so well (Mairesse 1974). It forces itself into relatively new areas such as frontier production functions (see Førsund et al. 1980). And it has been used both as a utility and production function in analyses of growth, development, macroeconomics, public finance, labour and just about any other applied area in economics. Yet it possesses restrictive properties and perhaps for that reason it has become for some an object of disdain, often regarded as a child's toy in the world of real economics. But for others, the Cobb–Douglas is at least a venerable form and, effectively, it and its putative inventor are regarded fondly.

In its unrestricted form, the Cobb–Douglas can be written as $f(x) = A\prod_{i=1} x_i^{ai}$, where A is an efficiency parameter, a_i is the elasticity of $f(x)$ with respect to x_i , and $\mathbf{x}$ is confined to $R^n{}_{++}$. Defining the x_i as goods consumed, it has been used as a utility function; defining them as inputs in the production process, it is a production function; as normalized prices, it is an indirect utility function; and so on. We focus here on its use as a production function for a single output.

A large part of the appeal of the form stems basically from the fact that if $0 < a_i < 1$, $f(\mathbf{x})$ is strongly pseudo-concave on its domain. That entails that if the firm is a profit maximizer and factor supply and product demand functions are continuously differentiable on their domains, then the input demand and supply of output functions have the immensely useful property of continuous differentiability everywhere on their respective domains. Also, if $\sum_i a_i \leq 1$ and if factor supply and product demand functions are well-behaved, the input demand functions are downward sloping with respect to own price and the output supply

function does not slope downward with regard to product price. What could be better and, moreover, it is all so simple to demonstrate.

Another attractive property of the form is that it has a function coefficient that is identical to its degree of homogeneity, calculated by summing the factor production elasticities. Thus, $\sum_i a_i \lessgtr 1$ for all i easily and succinctly characterizes decreasing, constant and increasing returns to scale, respectively. This characteristic also has important implications for the cost, profit and revenue duals of the production function. For example, the cost function of a price-taking firm which has a Cobb–Douglas technology decomposes into two parts, one a linear homogeneous function of factor prices and the other a function of output q, that is $C(q, \mathbf{w}) = B\prod_{i=1} w_i^{ci} q^{co}$, where B is a positive constant, $\mathbf{w}$ is a (positive) price vector of the inputs, $c_i = a_i/\sum_i a_i$ and $c_0 = 1/\sum_i a_i$.

The list of attractive properties extends to the aggregation problem since the Cobb–Douglas is homogeneous and weakly separable. First consider the question of aggregation across inputs. Suppose one can write a generalized Cobb–Douglas function as follows:

$$q = \prod_{s=1} \left(\prod_{j=1} x_{sj}^{b_{sj}} \right)^{Y_s},$$

where $b_{sj} = a_{sj}/\sum_j a_{sj}$, $Y_s = \sum_j a_{sj}$, J_s is the number of factors in the sth group, S is the number of groups, $s = 1, 2,..., S$, and $j = 1, 2,. . ., J_s$. Notice that $\sum_j b_{sj} = 1$. Since each expression in the parentheses is homogeneous of degree one for each s, the profit maximization procedure can be decomposed into two stages and there exist quantity and price indexes (call them x_s and W_s, respectively) such that the expenditure on the sth group is $W_s x_s$ for s = 1, 2, ..., S.

With respect to aggregation across firms, suppose the rth firm's production function were

$$q_r = A_r x_{1r}^{c_{1r}} x_{2r}^{c_{2r}}, ..., x_{nr}^{c_{nr}},$$

where $\sum_i c_{ir} = 1$ and $i = 1, 2, ..., R$. It is evident that the expansion paths for all firms are straight lines through their respective origins. Then under the extremely restrictive conditions that the expansion paths for each firm are parallel (i.e. if $c_{ir} = c_{it} = c_i$ for each i and for all r and t), and that the first order conditions are satisfied, the R functions consistently aggregate to

$$q = x_1^{c_1} x_2^{c_2}, ..., x_n^{c_n},$$

a nicely behaved aggregate production function.

There is another way to look at the aggregation-across-firms problem that involves the Cobb–Douglas function. Suppose that factors in each firm are used in fixed proportions with the Leontief coefficients being distributed across all firms according to a Pareto distribution. Then a surprising result by Houthakker (see Sato 1975) is that the aggregate production function of the industry is a Cobb–Douglas form.

Of course, there is a price for these desirable implications and most of it is owing to the fact that the Cobb–Douglas technology entails that the elasticity of substitution takes on the knife-edge value of unity. If there is no technological change, a unit substitution elasticity implies that the income shares of all factors of production remain constant in the face of changes in things that are deemed germane such as saving, the rate of growth of the economy and relative factor supplies. Only the state of the technology matters in this instance, a highly disputable outcome. When technological change is allowed to proceed in a Cobb–Douglas world, it is a fact that Hicks-, Solow- and Harrod-neutral technological change are equivalent, thus blurring these distinctions. Another implication of the unit substitution elasticity of the (linear homogeneous) Cobb–Douglas form is that, used in growth models, it guarantees the existence and stability of equilibrium growth, again obscuring an important problem in economics.

Furthermore, it is a fact that the Cobb–Douglas form requires that each factor of production be essential in the sense that no factor may be completely substituted for another. Hence the domain of the function must be confined to the set of strictly positive real numbers. This is not particularly disturbing for situations in which the factors can be taken to be large aggregates but it does limit the analysis in other contexts.

Technological change is represented in the Cobb–Douglas by changes in the efficiency parameter A which are Hicks neutral, by changes in the scale of the factor inputs which are factor augmenting and also Hicks neutral, and by changes in the elasticities of production which may be Hicks non-neutral. However, the unit elasticity of substitution is restrictive in still another way: it cannot represent a technological advance that results in a change in the ease of substitution among factors of production.

What is the form's provenance? It is generally attributed to Paul Douglas and although he gracefully acknowledged (Douglas 1967) that Wicksteed and Walras were cognizant of it, he neglected to add Wicksell's name to the list. Be that as it may, Douglas relates in his gentle comments that in 1927 he asked a professor of mathematics, Charles Cobb, to devise a formula that could be used to measure the comparative effect of each of two factors of production upon the total product to satisfy a linear log–log relationship in his input and output data. His work encountered a host of theoretical concerns (see Brown 1966 for a discussion) aside from the capital, output and labour measures for which he was faulted. But the production form remained in spite (or perhaps because) of its restrictive properties.

Subsequent work has demonstrated that the Cobb–Douglas is a special case of a variety of forms and approaches. The constant elasticity of substitution (CES) production function is perhaps the most well known of the forms that yield the Cobb–Douglas as a special case, either by using L'Hôpital's rule when the elasticity of substitution goes to unity or it can be derived from certain expressions used in deriving the CES function (see Brown and De Cani 1963). Parenthetically, the CES, itself, is known to mathematicians as a mean of order t [i.e. $\left(\sum_{i=1} a_i x_i^t\right)^{\frac{1}{t}}$ for $t \neq 0$] so that, if one takes the limit as $t \rightarrow 0$, of course, the Cobb–Douglas emerges. Also, it can be derived from the translog production form (Christensen et al. 1973) and many others, besides, by judiciously restricting certain parameters. A different approach to the derivation of the Cobb–Douglas form has been taken by P. Zarembka (1987), who specifies each variable as $z(\lambda) = (z^{\lambda} - 1)/\lambda$ for $\lambda \neq 0$ and $z(\lambda) = \ln z$ for $\lambda = 0$. Then, applying this transformation to the production function, we would have $z_0 = q$ and $z_i = x_i$ for all i. Thus, if the z_k ($k = 0, 1, ..., n$) are related linearly, the transformation turns out to be a useful procedure in econometrics to treat the general problem of functional form, an important special case of which is the Cobb–Douglas.

In sum, though it is restrictive and sometimes regarded as an economic toy, the Cobb–Douglas form is remarkably robust in a vast variety of applications and that it will endure is hardly in question.

See Also

- ▶ Capital Theory (Paradoxes)
- ▶ CES Production Function
- ▶ Douglas, Paul Howard (1892–1976)
- ▶ Production Functions

Bibliography

Brown, M. 1966. *On the theory and measurement of technological change*. Cambridge: Cambridge University Press.

Brown, M., and J. De Cani. 1963. Technological change and the distribution of income. *International Economic Review* 4: 289–309.

Christensen, L.R., D.W. Jorgenson, and L.J. Lau. 1973. Transcendental logarithmic production frontiers. *Review of Economics and Statistics* 55: 28–45.

Douglas, P.H. 1948. Are there laws of production? *American Economic Review* 38: 1–41.

Douglas, P.H. 1967. Comments on the Cobb–Douglas production function. In *The theory and empirical analysis of production*, edited by. M. Brown, National Bureau of Economic Research, Studies in Income and Wealth No. 31. New York: Columbia University Press.

Førsund, F.R., C.A.K. Lovell, and P. Schmidt. 1980. A survey of frontier production functions and of their relationship to efficiency measurement. *Journal of Econometrics* 13: 5–25.

Mairesse, J. 1974. *Comparison of production function estimates*. Paris: Institut National de la Statistique et des Etudes Economiques.

Sato, K. 1975. *Production functions and aggregation*. Amsterdam: North-Holland.

Wicksell, K. 1958. *Selected papers on economic theory*. ed. Erik Lindahl. Cambridge, MA: Harvard.

Zarembka, P. 1987. Transformation of variables in econometrics. In *The new Palgrave: A dictionary of economics*, ed. J. Eatwell, M. Milgate, and P. Newman, vol. 4. London: Macmillan.

Cobbett, William (1763–1835)

N. W. Thompson

William Cobbett was born, appropriately, at the 'Jolly Farmer' in Farnham, Surrey, in 1763. He was by turns soldier, clerk, teacher, journalist and political agitator but whether in his early literary career as an anti-Jacobin pamphleteer or later in his role as combative radical his voice was that of the small farmer threatened by the forces of economic and social change which characterized the early phases of Britain's industrial revolution.

Cobbett's acerbic castigation of its major theorists shows he had little time for political economy and he would certainly have greeted with a wry guffaw his inclusion in a dictionary of economic thought. Nevertheless it was the case that much of his writing in the *Twopenny Register* (1816–20) and works such as *Paper against Gold* (1815) was given over to a discussion of economic questions and in particular the economic difficulties that confronted Britain in the post-Napoleonic war period, and even in his *Rural Rides* (1830) few opportunities were lost of discoursing on matters economic.

For Cobbett material impoverishment was a consequence of the political decisions affecting taxation, the Debt, the convertibility of the currency etc., which emanated from a Parliament corrupted by the influence of tax-gatherers, 'Change Alley men, sinecurists, placemen, Jews and Borough-tyrants. Crucial here was the passage of the Bank Restriction Act of 1797, which in suspending cash payments by the Bank of England had created a 'Paper money system' that robbed the 'industrious' of their 'earnings' producing 'that monster in civil society, starvation in the midst of abundance'. This paper money system was seen in itself as a means of appropriating the product of labour but most importantly, with the return to cash payments in 1819, it created a situation where the nation was compelled to pay in an appreciated medium of exchange debts contracted in a depreciated paper currency. The industrious were therefore forced to pay in gold what had been contracted in paper. Thus both the inflation of the Revolutionary and Napoleonic wars and the deflation that followed were seen by Cobbett as redistributing wealth in favour of the idle generally and the fundholders and bankers in particular. Justice demanded, therefore, that there should be an equitable adjustment in the level of taxation to take account of the appreciation in the value of money. For this, parliamentary reform and an end to political corruption were necessary prerequisites.

Under the undifferentiated heading of 'The Thing', Cobbett sought to attack all those who jeopardized his vision of a stable, hierarchically structured, rural economy and society, dominated by the independent yeoman farmer and free from 'the all-devouring Jew and tax-eater'. Specifically, he condemned 'the Funding and Manufacturing and Commercial and Taxing System' for their tendency to draw 'wealth into great masses . . . [and] man also into great masses', like 'the Great Wen' [London], which both corrupted its inhabitants and impoverished the country as a whole.

Cobbett's was an anti-industrial, anti-commercial, anti-urban and anti-City political economy with its ideological roots in the political radicalism of the eighteenth century and with Paine's *Decline and Fall of the English System of Finance* (1797) as its basic text. Yet if, with the growth of industrial capitalism, this 'old corruption' critique became increasingly irrelevant and inadequate; if by the 1830s it was being supplanted in the working-class press by anti-capitalist and socialist analysis of contemporary ills which deployed to critical effect the tools and concepts of political economy, this should not obscure the remarkable longevity of Cobbettian ideas. Thus even a superficial survey of Chartist literature will show just how indelible was the mark left by the *Twopenny Register* upon a generation of political radicals. In a sense too Cobbett's vision of a pastoral England peopled

by free-born, hearty independent cultivators will remain indestructible, drawing sustenance as it does from an ineradicable human craving for the permanence, the certainty, the order and the stability which rural self-sufficiency seems to offer.

Selected Works

1802–36. *Cobbett's political register.* The paper was published throughout this period but the cheap, popular post-Napoleonic war edition of the paper referred to as *Cobbett's Twopenny Register* ran from 12 October 1816 until 29 July 1820.
1815. *Paper against Gold and Glory against prosperity.* London.
1830. *Rural rides in the counties of Surrey, Kent, Sussex*. . .. London.

References

Beer, M. 1953. *A history of British socialism*, vol. 2. London: Allen & Unwin.
Cole, G.D.H. 1947. *The life of William Cobbett*, 3rd ed. London: Home & van Thal.
Cole, G.D.H. 1977. *A history of socialist thought. 5 vols. Vol. 1, Socialist thought: The Forerunners, 1789–1850*. London: Macmillan.
Green, D. 1983. *Great Cobbett: The Noblest agitator*. Oxford: Oxford University Press.
Sambrook, J. 1973. *William Cobbett*. London: Routledge & Kegan Paul.
Spater, G. 1982. *William Cobbett: The poor man's friend*, vol. 2. Cambridge: Cambridge University Press.
Thompson, N.W. 1984. *The people's science: The popular political economy of exploitation and crisis*. Cambridge: Cambridge University Press.

Cobden, Richard (1804–1865)

William D. Grampp

Keywords

Anti-Corn Law League (UK); Bank of England; Bright, J.; Cobden, R.; Colonialism; Corns Laws; Factory Acts; Free trade; Manchester School; McCulloch, J. R.; New Poor Law; Pacifism; Tooke, T.

JEL Classifications

B31

Cobden led the campaign that repealed the Corn Laws in 1846, after which there was free trade in grain. The son of a Middlesex farmer, he sought his fortune in Manchester, became an owner of a mill that employed 2,000 workers and was noted for excellence of its calicoes. At 35, he was a rich man.

His calling, however, was politics. After taking part in the successful effort to incorporate Manchester, he entered the movement against the Corn Laws in 1838. Until then it had been conducted by middle class radicals and various business interests, among them the Manchester Chamber of Commerce. Cobden, John Bright, and others like them wanted to enlarge the movement, make it bold and uncompromising. They were exasperated by the businessmen who so wanted to look respectable that they could not see where their interest lay. Thomas Tooke had said the same about the London merchants, when on their behalf he drafted the celebrated petition of 1820 for free trade and they were reluctant to sign it.

The militants of Manchester formed the National Anti-Corn Law League and agitated for free trade up and down the country. They become known as the Manchester School of Economics and were celebrated as arch advocates of laissez-faire. Actually they were a coalition of diverse interests that agreed on only one issue – repeal of the Corn Laws – and each did so for its particular reasons.

Cobden's reason was peace. He believed free trade would break down national barriers and give everyone a material interest in avoiding war. This was not an argument gotten up for the occasion but the expression of a view he had long held. When young he wrote two long tracts on foreign policy which denounced alliances among nations and political engagements of all kinds, decried the idea of a balance of power, was especially disapproving of colonies, then went on to extol

free trade as the way to peace and its guarantor. Years later, after he and Bright had brought down the Corn Laws, he told him, 'I have always had an instinctive monomania against the system of foreign interference, protocolling, diplomatising, etc.'

That scarcely expressed the horror he had of violent action, even the suggestion of it. When the southern states of America seceded, he thought Lincoln was wrong in bringing the issue to battle although he had no sympathy with them (except their fondness for free trade). He was shocked by the massacres in India and was opposed to wars of independence and to revolution. He thought duelling was barbarous, was against capital punishment, objected to boxing, couldn't stand brass bands, and asked the Pope to prohibit bull fighting in Spain. He favoured free trade so long as its effect was peaceful, as he believed it usually was, but when he believed it was not he quickly put it aside. He opposed the sale of foreign bonds in the London market if the proceeds were to be used to buy arms. 'No free trade in cutting throats', he said.

Pacifism, not laissez-faire, was Cobden's guiding principle; and he applied laissez-faire less to domestic than to foreign markets. He did not care for the Factory Acts but only spoke, never voted, against them. He approved of increasing the monopoly powers of the Bank of England and of regulating aspects of railway construction. He had no use for the New Poor Law, of which most economists of the day approved, and spoke derisively of McCulloch's 'usual dogmatism'. But he carefully read the latter's edition of the *Wealth of Nations* and wrote in the margins of the chapters that moved him. His notes are especially lively where Smith condemns the colonial policy of Great Britain. However, where he describes the operation of the invisible hand, the margins are quite untouched.

See Also

► Corn Laws, Free Trade and Protectionism
► Manchester School

Selected Works

1835. *England, Ireland and America, by a Manchester Manufacturer.* London.
1836. *Russia, by a Manchester Manufacturer.* Edinburgh.
1849. *Speeches on Peace, Financial Reform and Other Subjects.* London.
1867. *Political Writings.* London: Ridgeway.
1868. *Speeches on Questions of Public Policy*, 2 vols, ed. J. Bright and J.E. Thorold Rogers. Oxford: Oxford University Press.

Bibliography

Ashworth, H. 1877. *Recollections of Richard Cobden, M.P., and the Anti-Corn Law League*. London: Cassell.
Morley, J. 1881. *The Life of Richard Cobden*. London: Fisher Unwin.
Ritchie, J.E. 1865. *The Life of Richard Cobden*. London: Ward and Lock.

Cobweb Theorem

B. Peter Pashigian

Abstract

The cobweb theorem purports to explain persistent fluctuations of prices in selected agricultural markets. It was first developed in the 1930s under static price expectations where the predicted price equalled actual price in the last period. Muth's rational expectations hypothesis posited that forecast errors will not be serially correlated and the pattern of past forecast errors cannot be used to improve the accuracy of the forecasts. The fundamental question of whether observed price cycles are better explained by systematic errors in price forecasts or by the cumulative impact of unpredictable shocks has not as yet been definitively addressed.

Keywords

Adaptive expectations; Cobweb theorem; Expectations; Kaldor, N.; Price cycles; Price expectations; Rational expectations

JEL Classifications

Q11

The persistent fluctuations of prices in selected agricultural markets have attracted the attention of economists from time to time, and the theory of the cobweb was developed to explain them. The theory is applicable to those markets where production takes time, where the quantity produced depends on the price anticipated at the time of sale, and where supply at time of sale determines the actual market price.

One strand of the cobweb literature (the term was coined by Kaldor 1934) concentrates on how expectations are formed and the effect of the price expectations mechanism on the stability of equilibrium. Cobweb theory was first developed under static price expectations where the predicted price equalled actual price in the last period. The cobweb theorem proved that the market price would (not) converge to (long-run) equilibrium price if the absolute value of the price elasticity of demand was greater (smaller) than the price elasticity of supply. This stability condition was modified later as more sophisticated expectations models were adopted. Early articles by Tinbergen, Ricci and Schultz appeared in 1930 in German (see Waugh 1964, for a review of this literature). Ezekiel's important article (1938) spells out in greater detail the conditions for convergence, divergence or perpetual oscillation and shows how cycles of different lengths could be generated under static expectations.

Why the theory was developed in the 1930s and not earlier is a bit of mystery, for recurring price cycles for some agricultural products had been reported by agricultural economists for some time. Economists may have been attracted to the cobweb theory in the 1930s because of the events of the Depression. A theory that explained both oscillation and long departures from stationary equilibrium was more attractive after the events of the Depression. The fact that Ezekiel's paper was reprinted in the 1944 American Economic Association volume on business cycles lends credence to this view.

The impression left by Ezekiel and subsequent contributors is that the cobweb theory is a valuable tool for explaining price cycles. Ezekiel was aware of the simplicity of static expectations and not unmindful of the importance of shocks on the demand and the supply sides of the market in causing aberrant price fluctuations (for example, weather and the randomness of yields). Even so, agriculture economists, who were presumably more familiar with price fluctuations in agricultural markets, have been more prone to accept the theory, while other theorists have given the theory more of a mixed reception.

The price expectations mechanism has undergone many refinements over the years. In 1958 Nerlove proposed the use of adaptive expectations. This suggestion is motivated by the findings of econometric studies which showed the price elasticity of demand to be less than the price elasticity of supply for many agricultural goods. Under these conditions the static expectations version of the cobweb model predicts a price cycle of increasing amplitude. However, the observed price cycles in agricultural markets showed no sign of being explosive. Nerlove attempted to reconcile theory with evidence and to show that convergence is possible under a broader set of conditions provided expectations are adaptive. During the 1930s the attractiveness of the cobweb model seemed to be in its ability to explain persistent or even explosive price cycles. By the late 1950s these were no longer attractive features, and Nerlove felt compelled to offer an explanation of why price cycles of increasing amplitude are not observed even when demand elasticities are smaller than supply elasticities. Waugh (1964) took a different tack and attempted to reconcile the theory with the evidence of stable price cycles by suggesting that the price elasticity of supply becomes smaller (larger) than the price elasticity of demand at prices well above (below) the long-run equilibrium price. Under this assumption, a stable price cycle will eventually be reached.

The length of the cobweb price cycle is determined by the length of the production process. If it takes one year to bring a fattened hug to the market, then the complete price cycle should take two years. At first, little attention and superficial explanations were given to explain why the predicted length is often shorter than the actual length of the price cycle. It was left to the critics to point out these discrepancies.

The critics are responsible for the other strand of the literature. They appeared early but were not very influential at first although their criticisms were ultimately given more weight. The critics questioned the rationality of using an arbitrary expectations mechanism by otherwise profit-maximizing agents, and pointed out that the theory implies that producers would expect to lose wealth if they entered and remained in an industry with a cobweb price cycle. In a perceptive article on the pig cycle in England, Coase and Fowler (1935) questioned the realism of static expectations. They showed that the price of a bacon (mature) pig less the cost of feeding for the next five months and less the cost of a feeder (*y*oung) pig, which would be stable in a competitive market if farmers had static expectations, fluctuated over time. Hence the empirical evidence contradicted the assumption of static expectations. They presented evidence that pig breeders reacted quickly to a change in expected profits, and this implied that the pig price cycle should be only two years instead of the observed four-year period. The fluctuation in the profits per pig was attributed to the difficulty of predicting both demand and foreign imports. The Coase-Fowler paper advanced, if only in faint outline, the essence of the rational expectations hypothesis which was to blossom some 35 years later. They hinted that anticipated prices would not be formed in a mechanistic way because profits would be higher the more accurate are the forecasts. Prediction errors were due to the difficulty of predicting shifts in demand and in foreign supply.

Buchanan's paper (1939) criticized the cobweb model because it implied that producers suffer aggregate losses over the price cycle when output is determined by the long-run supply curve. He pointed out that the theory was based on the dubious assumption of a continued supply of entrepreneurs standing ready to dissipate their capital. The critics were also disturbed by the ambiguity of whether the supply curve is of the short-or long-run variety, and the failure to clarify how the adjustment from the short-run to the long-run supply curve is made. These early criticisms and ambiguities aside, references to the cobweb theory continued to appear in textbooks.

Nerlove's paper (1958) briefly rekindled the controversy. His purpose was to resurrect the theory and show that it could explain price behaviour if adaptive expectations were employed. Mills (1961) criticized the use of adaptive and other autoregressive expectations mechanisms in the deterministic model because they implied a simple pattern of forecast errors that producers could detect, incorporate into their forecasts and thereby improve the accuracy of their price forecasts. While Nerlove's suggestion did rectify one limitation of the cobweb theory, it did not address the critical issue of why producers relied on any particular forecasting mechanism. Muth (1961) developed the implications of rational expectations for cobweb theory in his now famous paper. Muth postulated that expectations were the predictions of the economic structure of the market and incorporated all available information. Under certain conditions the predicted price equals the conditional expectation of price, given currently available information. Adaptive expectations can be rational only under special conditions, and the coefficient of adaptation is determined by the values of the slopes of the demand and supply curves.

The rational expectations formulation has powerful implications for cobweb theory. If the price forecasts incorporate all available information and are on average correct, then forecast errors will not be serially correlated and the pattern of past forecast errors cannot be used to improve the accuracy of the forecasts. Moreover, what is then left of the supposed ability of the cobweb theory to explain the cyclical behaviour of prices? Price fluctuations would have to be explained either by the cyclical pattern of exogenous variables or by the summation of random shocks (Slutsky 1937). Muth's paper represents a

frontal attack on the traditional cobweb model. He notes that the traditional model tends to predict a shorter price cycle than is observed and indicates that the rational expectations version predicts a longer price cycle.

Interest in the cobweb model has ebbed in recent years and few articles on it have appeared in the major journals. Economists have found it more rewarding to apply the rational expectations hypothesis to areas like monetary or business-cycle theory than to the study of particular markets, even though the analysis of markets with inventories raises issues that are just as difficult and subtle. The question of whether the cobweb does or does not explain price cycles has not really been resolved. Freeman (1971) has suggested that the traditional cobweb model explains cycles in the markets for lawyers, physicists and engineers. Tests of the rational expectations hypothesis have been suggested by Pashigian (1970) when expectations data are available and by Hoffman and Schmidt (1981) when expectations data are unavailable. So the methodology exists for distinguishing between the competing hypotheses. Few econometric tests have been made of the rational expectations hypothesis in markets where the assumptions of the cobweb model apply. The fundamental question of whether observed price cycles are better explained by systematic errors in price forecasts or by the cumulative impact of unpredictable shocks has not as yet been definitively addressed.

See Also

- ▶ Adaptive Expectations
- ▶ Rational Expectations

Bibliography

Buchanan, N. 1939. A reconsideration of the cobweb theorem. *Journal of Political Economy* 47: 67–81.

Coase, R.H., and R.F. Fowler. 1935. Bacon production and the pig-cycle in Great Britain. *Economica* 2: 142–167.

Ezekiel, M. 1938. The cobweb theorem. *Quarterly Journal of Economics* 52: 225–280.

Freeman, R.H. 1971. *The market for College-trained manpower*. Cambridge, MA: Harvard University Press.

Hoffman, D.L., and P. Schmidt. 1981. Testing the restrictions implied by the rational expectations hypothesis. *Journal of Econometrics* 15: 265–287.

Kaldor, N. 1934. A classificatory note on the determinateness of equilibrium. *Review of Economic Studies* 1: 122–136.

Mills, E.S. 1961. The use of adaptive expectations in stability analysis: Comment. *Quarterly Journal of Economics* 75: 330–335.

Muth, J.F. 1961. Rational expectations and the theory of price movements. *Econometrics* 29: 315–335.

Nerlove, M. 1958. Adaptive expectations and cobweb phenomena. *Quarterly Journal of Economics* 72: 227–240.

Pashigian, B.P. 1970. Rational expectations and the cobweb theory. *Journal of Political Economy* 78: 338–352.

Slutzky, E.E. 1937. The summation of random causes as the source of cyclical processes. *Econometrica* 5: 105–146.

Waugh, F.V. 1964. Cobweb models. *Journal of Farm Economics* 46: 732–750.

Coddington, Alan (1941–1982)

Victoria Chick

Born in Doncaster, Yorkshire, Coddington began his academic career with a degree in physics at Leeds University. After a year teaching mathematics in a school in York, he returned to university in that city for his D.Phil. in Economics. On taking his degree in 1966 he was appointed Assistant Lecturer in Economics at Queen Mary College, London, where he rose steadily to become Professor in 1980.

His theoretical work has three main strands: an early interest in the theory of bargaining, resulting in several articles and a much-respected book (1968); various aspects of environmental economics; and, continuing throughout his career, methodology and the history of 20th-century economic thought.

It is this last area which one immediately associates with Coddington's name. From his work on interpretations of Keynes, his characterizations of the 'neoclassical synthesis' of the textbooks as 'hydraulic Keynesianism' and of the 'fundamentalists' as 'Chapter 12 Keynesians' (referring to the chapter of the *General Theory* which discusses the precariousness of long-term expectations) have entered economists' everyday language.

His coverage of modern economic thought included not only Keynes, but also Hicks, Shackle, Friedman, Hahn, Malinvaud, Clower and Leijonhufvud. His article on Hicks (1979) and much of his work of Keynes, with an invaluable new chapter ('The Keynesian Dichotomy') were reworked into his posthumously published book, *Keynesian Economics: The Search for First Principles*, to which every reviewer responded first with enthusiasm, then with regret that they had been deprived by Coddington's suicide of the intellectual stimulus of debating with him.

Selected Works

1968. *Theories of the bargaining process*. London: George Allen & Unwin.
1973. Bargaining as a decision process. *Swedish Journal of Economics* 75(4): 397–405.
1974a. The economics of conservation. In *Conservation in practice*, ed. A. Warren and F.B. Goldsmith. New York: Wiley.
1974b. Creaking semaphore and beyond: A consideration of Shackle's epistemics and economics. *British Journal for the Philosophy of Science* 26(2): 151–63.
1975. The rationale of general equilibrium theory. *Economic Inquiry* 13(4): 539–58.
1978. Review of the theory of unemployment reconsidered (Malinvaud). *Journal of Economic Literature* 16(3): 1012–18.
1979. Hicks' contribution to Keynesian economics. *Journal of Economic Literature* 17(3): 970–88.
1983. *Keynesian economics: The search for first principles*. London: George Allen & Unwin.

Codetermination and Profit-Sharing

D. M. Nuti

The contract regulating labour employment by capitalist firms usually embodies three basic elements: a fixed money wage rate per unit of time, the subjection of workers to the employer's authority in the workplace and the short-term nature of the hiring commitment. Explicit or implicit departures from this standard can be observed; they are the result of individual or collective negotiations in the labour market, which balance out their advantages and disadvantages for each party, either directly or through accompanying changes in other parameters of the labour contract. Government legislation and economic policy set limits or fix actual values for some of these parameters and stipulations; within these bounds the market determines the rest.

Long tenure, i.e. the employee's option on continued employment, like all options has a value (for the employee) and a cost (for the employer), which is matched by correspondingly lower pay than that associated with shorter-term contracts. The partial and delayed indexation of money wages to a consumer price index for the period between successive rounds of wage negotiations favours employees when inflation decelerates and employers when it accelerates. Piece-rates, i.e. wages related to *individual* performance, give employees a short term reward (penalty) for effort supply higher (lower) than that which otherwise could be contractually fixed, as well as automatic participation in productivity gains due to learning by doing, subject to a ratchet effect on the determination of subsequent rates; employers save on the costs of recruitment, supervision, and contractual enforcement, lose short term productivity gains but can use more fully their contractual power in exacting effort and speeding up progress when rates are reviewed. Government policy influences directly or indirectly market

choice, in the pursuit of policy targets such as distributive fairness, employment, price stability, efficiency and growth.

The same combination of private interest and government policy determines the degree of workers' participation in decision-making processes (codetermination) and in the performance (profit-sharing) of enterprises (for a bibliographical survey, see Bartlett and Uvalic 1985).

Codetermination

Employee participation in enterprise decision-making in cooperatives amounts to full entrepreneurship through participation in assemblies, the election of representative organs and involvement in the appointment of managers. In other enterprises it takes the form of access to information and right to consultation, participation in decisions on conditions and organization of work and on internal social questions, through a workers' council or similar organ; right up to the minority (or even parity) participation and vote in the board of directors of a joint-stock company (as in German *Mitbestimmung*; see Nutzinger 1983) with a possibility of influencing decisions about employment, the level and structure of investment and other crucial factors were the other board members to be sufficiently divided.

The effects of codetermination are three-fold:

(i) The reduction in labour disutility obtainable when workers have a say in the division of labour and work organization, since enterprises may neglect workers' preferences about the specific uses to which their labour is put or at any rate respond to the needs of a hypothetical average worker: if the number of enterprises is not large enough, workers' control is necessary to reduce disutility and alienation. The effect of workers' control on productivity has an indeterminate sign (Pagano 1984).

(ii) The reduction of the number and intensity of conflicts in the workplace in general and, in particular, the more likely acceptance by workers of unpopular decisions by management, when workers receive detailed and credible information and participate in decision making, identifying themselves partly with the enterprise and above all lengthening their time horizon in view of continued participation in decision-making (Aoki 1984; Cable 1984; Fitzroy and Mueller 1984). Of course conflicts within the firm are made more tractable by the *introduction* of codetermination but *afterwards* are bound to reappear over time (Furobotn 1985); also there remains a basic conflict between employed and unemployed workers which may even be exacerbated by the employment protection policies conceivably encouraged by those already employed in their exercise of codetermination.

(iii) The greater correspondence between workers' powers and responsibilities, codetermination being the counterpart of workers' exposure to enterprise risks. The very fact that workers, unlike capitalists, cannot diversify between different enterprises when selling their services exposes them to an employment and income risk which induces them to make a claim to control; a claim which up to a point the employer may prefer to accept instead of granting higher wages or longer tenure.

Profit-Sharing

In pre-capitalist systems workers' participation in the results of their enterprise took the forms – now little used – of sharecropping in agriculture and of sliding scales (indexing wage rates to the price of the product), for instance, in English coalmines. In modern capitalism such participation – for which 'profit-sharing' is a shorthand label – takes the form of cooperatives' net revenue sharing, production prizes based on group or overall performance, participation in gross/net revenue/profit, share options, participation in investment funds and pay increases graded according to productivity growth.

The effects of an element of profit-sharing in labour earnings are three-fold:

(i) An expected increase in labour productivity. This is not due to workers gaining from the product of *individual* extra-effort (as in the case of piece-rates) since each of n workers employed will only get at most $1/n$ of the product of his own extra-effort (Samuelson 1977) and on the contrary may *reduce* effort if he can, being exposed to at most only $1/n$ of the output loss from his own lower effort. The productivity gain can be expected from workers, costlessly to themselves, gaining from intelligent and effective use of any given individual level of effort, from cooperating with other workers and management and from monitoring and supervising each other's effort, efficiency and cooperation (Reich and Devine 1981; Fitzroy and Kraft 1985).

(ii) Cyclical flexibility of labour earnings and therefore greater stability of profit levels and rates. Employment will not be stabilized during the cycle by labour earnings flexibility obtained through profit-sharing because the marginal cost of labour to firms – i.e. the fixed component of pay – does not vary automatically. Workers, who are normally risk-averse, will prefer a fixed sum of money to a profit-sharing formula of equivalent amount while employers, who are normally risk-lovers, may or may not prefer greater stability of profit rates (according to their actual attitude to risk and the alternative cost of reducing risk through diversification) to the point of granting higher average earnings on a profit-sharing formula than a fixed wage to mutual advantage. Therefore profit-sharing is favoured primarily in risky ventures; otherwise on this ground alone profit-sharing would be favoured by firms only in a recession (when workers would only accept it as an alternative to a permanent wage cut) and by workers only during a boom (when firms would only accept it as an alternative to a permanent wage increase).

(iii) Higher level of labour employment, for a given level of labour earnings with respect to a fixed wage regime, due to the lower marginal cost of labour to profit-sharing firms. Vanek finds that higher employment will be associated with higher aggregate income, lower prices (because of higher output), higher export volume and domestic import substitution (with undetermined effects on the balance of payments depending on price and income elasticities), lower after-tax and after-labour-share profits and higher labour-share in national income (Vanek 1965).

Rediscovering Vanek's macroeconomic benefits from profit-sharing (though not its impact on net profits and relative income shares), Weitzman claims that these benefits are neglected by individual firms, as in other instances of 'public goods', 'externalities' and 'market failures', therefore necessitating public policy measures (Weitzman 1983, 1984). However, there is no reason why a firm should object to granting a given increase in earnings under the guise of a profit-share instead of an equivalent fixed amount unless that represents forced insurance against profit variability; and why workers – at least at the level of nation-wide collective bargaining – should not take into account the potential employment and price stability benefits of this formula and offset them against the greater variability of their earnings in between negotiations, due to both cyclical factors and random factors affecting their firm's performance.

Contrary to Weitzman's belief, in fact, profit-sharing is not absolutely superior to wage contracts. For workers, profit-sharing transforms the probability distribution of uncertain employment at a fixed and certain income into a probability distribution of employment with a higher mean (because of lower marginal cost of labour) but no less variable over the cycle, at a more variable income (both over the cycle and for other factors affecting dispersion of enterprise performance) and at a higher (real) mean. For firms it transforms a more into a less variable probability distribution of money profit rates around the same mean (or a lower mean if workers are protected from actual losses; the effect on real profit rates depending on accounting conventions and choice of *numéraire*). In the pursuit of greater employment and price

stability of course a government may grant tax relief to shared profits, just as effectively and with just as much reason as it may subsidize the marginal cost of labour to firms under a wage regime. Otherwise there is no reason why profit-sharing should be forced upon unwilling workers and firms by well-meaning reformers, beyond the extent they are prepared to consider in their market transactions. These propositions are developed further below (see also Nuti 1985, 1986).

Interdependence Between Codetermination and Profit-Sharing

The respective effects of codetermination and of profit-sharing are not independent. The productivity increase expected from profit-sharing can be raised by workers having collective discretion over the organization of labour; or the productivity fall which might derive from workers' control over labour organization might be tempered by profit-sharing. Greater variability of earnings – during the cycle and across firms – strengthens under profit-sharing the case for codetermination already present in workers' exposure to employment risk in the wage régime. The income premium required by risk-averse workers to replace some of their fixed wage with a variable profit-share can be reduced by their involvement in the decisions which expose them to income variability in the first place. The reduction in conflict frequency and intensity expected from codetermination is enhanced by profit-sharing because for each worker it partly internalizes the conflict between 'us' and 'them' otherwise manifested and enacted externally; in any case it is a requirement of any effective incentive system that power and responsibility should not be separated.

The quantification of degrees of 'codetermination' and to a lesser extent of 'profit-sharing' raises conceptual and practical difficulties (though see Cable 1985). By and large we can observe a certain correlation between the two: both codetermination and profit-sharing are zero in the pure capitalist enterprises and unity in cooperatives and other forms of partnerships of capital and labour; minor forms of codetermination (or conversely of profit-sharing) tend to go hand in hand with minor forms of profit-sharing (or of codetermination); a high degree of one without the other is virtually unknown.

The combination of 100 per cent codetermination (= self-determination) and 100 per cent profit-sharing (= net revenue sharing) obtained in cooperative firms, according to conventional literature, is subject to economic stimuli of a somewhat 'perverse' kind. These are primarily: restrictive employment (= membership) policies; destabilizing and Pareto-inefficient reactions (or at best inelasticities) to price changes and technical progress; a low propensity towards self-financed investment (Ward 1958; Vanek 1970). In empirical studies of cooperative firms there is no incontrovertible evidence of these phenomena, which are probably partly offset by other economic (job security, growth-mindedness, etc.) and non-economic stimuli; but there is a presumption that – albeit in a weak form – the same tendencies and, in particular, employment restrictive policies might be associated with codetermination. We can also presume that workers' eagerness to press and ability to assert demands for codetermination, as in the case of other demands, increase as unemployment diminishes. Hence the employment-generating benefits of profit-sharing can be at least partly offset by the restrictive employment policies possibly associated with codetermination brought about by profit-sharing and by greater proximity to full employment. Recent empirical studies suggest modest but sizeable improvements in economic performance from codetermination and profit-sharing (Cable and Fitzroy 1980; Estrin et al. 1984) when and where they occur but there may have been costs that remained unobserved and, in any case, the improvements cannot be generalized.

Markets and Policy

Degrees of codetermination and profit-sharing may well be regarded as desirable on 'political' (as opposed to 'purely technical') grounds such as

equity and social peace. They may also be the best policy instruments in the pursuit of public objectives such as stability, employment and growth, in the sense of having the least cost in terms of public funds or offering the most attractive trade-offs between alternative targets. Otherwise, as Jensen and Meckling argue for codetermination and one can also argue for profit-sharing, if it is truly beneficial to both stockholders and labour no laws would be needed to force firms to undertake reorganization (1979, p. 474). Yet renewed and insistent calls for public intervention in favour of *profit-sharing without codetermination* have been put forward by M.L. Weitzman in recent writings (1983, 1984, 1985a, b, 1986). The proposal has been enthusiastically received in certain academic and political circles and hailed as a breakthrough in the specialist press.

Weitzman's novelty, the foundation for this renewed fascination with profit-sharing, is the rash assertion of two propositions. First, that long-run full employment equilibrium under profitsharing is associated with permanent but non-inflationary excess demand for labour, which cushions off the economy from contractionary shocks and gives new dignity and status to labour. In adman's language we are told, for instance:

> A share system has the hard-boiled property of excess demand for labour, which turns into a tenacious natural enemy of stagnation and inflation. The share economy possesses a built-in, three-pronged assault on unemployment, stagnant output, and the tendency of prices to rise. This is a hard combination to beat. (Weitzman 1984, p. 144.

Second, that even in the short run the share economy can achieve and maintain full employment. For instance:

> The share system, . . ., has a strong built-in mechanism that automatically stabilizes the economy at full employment, even before the long-run tendencies have had the chance to assert their dominance. . . . a share economy has the direct 'strong force' of positive excess demand for labor . . . pulling it towards full employment. . . . the strong force of the share system will maintain full employment. (Weitzman 1984, p. 97)

Were these claims well founded an enlightened government possessing these truths would be justified in forcing profit-sharing on to a yet unconverted and disbelieving public, thus achieving full employment, price stability and growth at a stroke. Unfortunately miracles exist only for the uninformed and the faithful, but do not bear the weight of sober scrutiny. First, excess demand for labour at full employment cannot be sustained and can only be a temporary disequilibrium. Second, permanent excess demand for labour is inconsistent with lack of codetermination, and when this is introduced restrictive employment policies will alter the picture. Third, and most important, there is no guarantee that full employment can necessarily be achieved. Without these benefits the alleged 'public good' merits of the sharing contract disappear.

Excess Demand for Labour at Full Employment

Suppose that the share economy reaches a state of full employment. Weitzman maintains the presence and persistence of excess demand for labour in long-run equilibrium on the basis of the following argument:

$$\text{labour total pay} = \text{marginal revenue value of labour productivity at full employment} \tag{1}$$

because long-run equilibrium must be full-employment equilibrium and because of the underlying homomorphism of profit-sharing and wage contracts in long-run equilibrium (Weitzman 1983). By definition of profit-sharing

$$\text{labour total pay} = \text{fixed pay} + \text{share of net profits} \tag{2}$$

where fixed pay is greater than or equal to zero, and the share of net profits is greater than zero. It follows from (1) and (2) that:

$$\text{marginal revenue value of labour productivity at full employment} > \text{fixed pay} = \text{marginal cost of labour to firms} \quad (3)$$

i.e. firms will wish to employ more workers than are available. A permanent state of excess demand for labour will exist, which will protect full employment from contractionary shocks, as long as shocks do not reduce the marginal revenue value of labour productivity at full employment below the fixed element of pay, in which case the maintenance of over-full employment requires a reduction of the fixed element without cutting earnings as much as necessary in the wage regime.

There are three grounds for refuting this syllogism. First, firms should be well aware that, whatever their pay formula, they can only attract workers by offering the going rate for labour total pay and should regard this, and not the fixed element of pay, as marginal cost of labour. If firms behave as they should, excess demand for labour disappears.

Second, if firms regard the fixed element of pay as the marginal cost of labour they should find its being lower than the marginal revenue value of labour productivity disquieting enough to experiment with alternative combinations of pay parameters without raising total pay above labour productivity. Risk-averse workers preferring fixed pay to potentially variable earnings of identical mean, risk-neutral or risk-loving employers will reduce their labour cost by raising the fixed element of pay at the expense of workers' profit share; even without taking into account attitude to risk it is plausible to expect managers to experiment with alternative pay parameters and not to rest until they have equalized their marginal cost and marginal value of labour, i.e.

$$\text{marginal revenue value of labour productivity at full employment} = \text{fixed pay} \quad (3')$$

which can only be reconciled with the definition (2) of a profit-sharing contract if the workers' share of net profit is zero: with the sharing component of earnings the 'share economy' also vanishes and reverts to the fixed wage economy without any excess demand for labour.

Third, workers perceiving excess demand for labour are likely to reduce their supply of effort and/or increase turnover – as they do in the only known instances of permanent excess demand for labour, i.e. Soviet-type economies (see Lane 1985) – if not right down to the point where their marginal product equals fixed pay at least as close to that level as they are allowed to get by monitoring and supervising arrangements. This is another mechanism which can reduce and eliminate excess demand for labour if it occurred.

Codetermination and Employment

The lack of codetermination is an explicit precondition of Weitzman's claims (though not of Vanek's, who does not claim full and over-full employment of labour and does not need this restriction). (In the earlier version of his analysis Weitzman takes a sanguine view of the possibility of keeping codetermination in check: '... the bargaining power of labor unions is not a natural right ...' (1984a, p. 109); '... the decisions on output, employment and pricing are essentially made by capitalists' in his model (p. 132); 'I can see no *compelling* reason why a capitalist firm should be more prone to allow increased worker participation in company decision making under one contract form than under another' (p. 133, emphasis added). His latest version is more open-minded: workers' participation in decision-making becomes not only possible but desirable as 'a question of justice and practical politics' as long as it excludes *employment* decisions (1986). It is extremely hard to imagine *any* major decision, in which workers might have a voice, that would *not* directly or indirectly also affect employment. Either this limitation or workers' participation would have to give way.) We know that it is possible to exclude workers from codetermination in the presence of persistent unemployment; such exclusion might be difficult at full employment, and it would certainly be very

difficult with excess demand for labour, but the *persistent* state of excess demand for labour postulated by Weitzman should make the exclusion of codetermination, whether or not employment questions are directly involved, impossible without an authoritarian or military regime. This is not a moral, or legal, or legalistic proposition; it is a question of 'practical politics'.

Once workers have a say on output, employment and pricing and related questions (investment, innovation, etc.) they will try and resist the very possibility of dilution of their own shares just as shareholders usually resist the dilution of share capital; for better or worse they are likely to adopt, or are tempted to adopt, other things being equal, restrictive employment policies in the possibly misguided and self-defeating purpose of raising or maintaining individual earnings. This is not a case *against* profit-sharing, but an argument for not expecting that overfull employment, if achievable, can be sustained necessarily, i.e. an argument against the plausibility of Weitzman's model (see Nuti 1985).

Profit-Sharing and Full Employment

The foundation of Weitzman's claims on behalf of profit-sharing is the assertion that, even in the short run, the share economy 'delivers' full employment of labour. ('Resources are always fully utilised in a share system' (Weitzman 1985b, p. 949); real world frictions, inertias and imperfections are mentioned only to be exorcised, and to reassert the full employment claim at least as a 'natural tendency' (p. 949, p. 952) of the share economy which, we are told, 'delivers full employment' (1986); see also Weitzman 1984, p. 97.)

For a share economy to 'deliver' full employment three necessary conditions must be satisfied simultaneously:

(i) The physical marginal productivity of labour at full employment must be positive;
(ii) The marginal revenue obtained by firms from that physical marginal product of labour must also be positive;
(iii) The fixed element of pay in share contracts must be flexible enough to fall down to the level of the marginal revenue product of labour at full employment, positive as it may be.

The first condition rules out the possibility of *classical* unemployment, i.e. due to lack of equipment, land or other resources in the quantities necessary to employ all workers efficiently. Yet, after over a decade of deep and protracted recession, deindustrialization and decapitalization, even advanced industrialized countries such as Britain of France today cannot be expected to be able to satisfy this condition as a matter of course, not to speak of Italy or, say, Spain, or of less developed countries. In his formal model Weitzman (1985b) postulates constant physical productivity of labour; this is a plausible assumption *up to near-full capacity* but Weitzman gives no reason why the capacity should be constrained by labour instead of other resources.

The second condition rules out the possibility of *Keynesian* unemployment, i.e. aggregate demand constraints making the marginal product of labour valueless before full employment is reached. Even if the first condition was satisfied, imperfect competition – which in all of Weitzman's work provides the environment in which the share contract is to operate – provides an excellent reason why firms might not give to additional physical products a positive value. Weitzman can assert that '... a "pure" sharing system not having any base wage would possess an infinite demand for labor' (1985b, p. 944), which implies positive marginal revenue for any level of output, because of the very special assumption that the elasticity of demand is *greater than unity* (p. 938), which makes demand curves absurdly and indefinitely elastic even for imperfectly competitive firms. The proposition cannot have any claim to general validity.

Even if demand for labour *were* to be infinite in the pure share economy, i.e. with a zero fixed element of pay, it would not necessarily be infinite, or even large enough to reach full employment, for a positive fixed element of pay. Weitzman neglects the determination of the

relative weight of the fixed and variable components of the share contract but recognizes the impossibility of total dependence of pay on profit; yet he takes for granted, for no good reason, that the fixed element of pay can be compressed down to whatever is the full employment marginal revenue product of labour, which we do not even know for sure is positive.

It is a non-controversial feature of the sharing contract, known from Vanek (1965), that the replacement of part of the wage by a profit-share of identical average cost to firms will lead to greater employment, higher output and lower prices – in the absence of large enough adverse feedback on investment (which Weitzman recognizes as a possible short run effect of the introduction of sharing) and in the absence of large enough feedbacks of accompanying codetermination on firms' employment policy. But there is a world of difference between higher employment and full employment and another world of difference between full employment and persistent over-full employment; no serious work can afford to switch indifferently and cavalierly from one to the other.

Share Contracts and Public Good

If the share economy could really guarantee, as general and necessary consequences of its establishment, the achievement and stability of full employment without adverse drawbacks there would be a case for public policy treating the share contract as 'public good' to be pressed on an unenlightened public still largely unaware of potential benefits, as in the case of safe vaccination against infectious disease. The case for the share economy would not be much greater than that for enforced wage flexibility, which would also guarantee full employment and stability under the same circumstances. A downward flexible wage would not deliver excess demand for labour but this is a questionable achievement and would not be necessary to absorb contractionary shocks if wages were flexible; downward flexible wages would also require a greater fall of money earnings to achieve full employment in the short run and may be more likely to bring about adverse effects on aggregate demand; otherwise there is little to choose between the two, except for the lower degree of public resistance that can be expected for share contracts with respect to wage cuts.

In fact if the share contract could really deliver and maintain full employment, while a wage economy could not, the greater variability of workers' earnings associated with profit-sharing over the cycle would disappear and, between firms, could be eliminated by labour freely redeploying itself at will across labour-hungry firms; the variability of employment would also disappear; workers would have de facto free access to a job in any firm of their choice, as in forgotten utopias (Hertzka 1890; Chilosi 1986). Thus it could be said that '... a move towards profit sharing represents an unambiguous improvement for the working class' (Weitzman 1985b, p. 954). But we have seen that profit-sharing cannot guarantee full (let alone over-full) employment. Without full employment, the higher variability of earnings associated with profit-sharing remains and it may or may not be compensated by the higher mean value of employment probability and perhaps real earnings. Outside over-full employment, in fact, the share economy is just as vulnerable to contractionary shocks as the wage economy because, in spite of flexibility of labour earnings in the share regime, the marginal cost of labour to firms (which is the fixed component of workers' pay) remains constant just as does the wage. Thus the higher stability of employment to be found in Japan simply cannot be the result of profit-sharing, as Weitzman firmly believes, seeing that Japan has never known a state of over-full employment; higher employment stability would require workers' shares in GNP instead of their enterprise's profits.

The fact that the adoption of a share contract, without the guarantee of stable full employment, has a cost for workers, eliminates the necessity, but not the possibility, of the share contract having 'public good' features. A vaccine may be somewhat unsafe, its degree of unsafety acceptable to all if vaccination is universal and all benefit from reduced exposure to infection, yet individuals benefit from free-riding strategies and the

enforcement of universal vaccination as 'public good' can still be beneficial to all. If labour contracts were negotiated exclusively at the level of individuals or firms the external beneficial effects of the share contract might be lost from sight; but these external benefits – unlike the case of genuine 'public goods' – are completely internalized in nationwide negotiations between associations of employers and employees. Admittedly the benefits, such as they are, of profit-sharing may be still unknown to the public at large and deserve wider publicity. But it is counterproductive to foist a good medicine on a sceptical public by claiming that it can guarantee longevity or immortality. At the first signs that such excessive claims are unfounded it may be thrown away despite its real lesser benefits.

Bibliography

Aoki, M. 1984. *The co-operative game theory of the firm*. Oxford: Oxford University Press.

Bartlett, W., and M. Uvalic. 1985. *Bibliography on labour-managed firms and employee participation*. European University Institute Working Paper, No. 85/198, Florence.

Cable, J.R. 1984. *Employee participation and firm performance: A prisoners' dilemma framework*. European University Institute Working Paper, No. 84/126, Florence.

Cable, J.R., and F.R. Fitzroy. 1980. Productive efficiency, incentives and employee participation: Some preliminary results for west Germany. *Kyklos* 33(2): 100–121.

Chilosi, A. 1986. *The right to employment principle and self-managed market socialism: A historical account and an analytical appraisal of some old ideas*. European University Institute Working Paper, No. 86/214, Florence.

Estrin, S., D.C. Jones, and J. Svejnar. 1984. *The varying nature, importance and productivity effects of worker participation: Evidence for contemporary producer cooperatives in industrialised Western societies*. CIRIEC Working Paper, No. 84/04, University of Liège.

Fitzroy, F.R. and K. Kraft. 1985. Profitability and profit-sharing. *Discussion Papers of the International Institute of Management*, WZB, Berlin, IIM/IP 85–41, December.

Fitzroy, F.R., and D.C. Mueller. 1984. Cooperation and conflict in contractual organisations. *Quarterly Review of Economics and Business* 24(4): 24–49.

Furobotn, E.G. 1985. Codetermination, productivity gains and the economics of the firm. *Oxford Economic Papers* 37: 22–39.

Hertzka, T. 1980. *Freiland. Ein soziales Zukunftsbild*. Dresden: Pierson. English translation, London: Chatto & Windus, 1891.

Jensen, M.C., and W.H. Meckling. 1979. Rights and production functions: An application to labor-managed firms and codetermination. *Journal of Business* 52: 469–506.

Lane, D. (ed.). 1985. *Employment and labour in the USSR*. London: Harvester Press.

Nuti, D.M. 1985. *The share economy: Plausibility and viability of Weitzman's model*. European University Institute Working Paper, No. 85/194, Florence; Italian translation in *Politica ed Economia* 1, January 1986.

Nuti, D.M. 1986. A rejoinder to Weitzman. (In Italian.) *Politica ed Economia* 4, April.

Nutzinger, H.G. 1983. Empirical research into German codetermination: Problems and perspectives. *Economic Analysis and Workers' Management* 17(4): 361–382.

Pagano, U. 1984. *Welfare, productivity and self-management*. European University Institute Working Paper, No. 84/128, Florence.

Reich, M., and J. Devine. 1981. The microeconomics of conflict and hierarchy in capitalist production. *Review of Radical Political Economics* 12(4): 27–45.

Samuelson, P.A. 1977. Thoughts on profit-sharing. *Zeitschrift für die Gesamte Staatswissenschaft* (special issue on profit-sharing).

Vanek, J. 1965. Workers' profit participation, unemployment and the Keynesian equilibrium. *Weltwirtschaftliches Archiv* 94(2): 206–214.

Vanek, J. 1970. *The general theory of labor-managed market economies*. Ithaca: Cornell University Press.

Ward, B.M. 1958. The firm in Illyria: Market syndicalism. *American Economic Review* 48(4): 566–589.

Weitzman, M.L. 1983. Some macroeconomic implications of alternative compensation systems. *Economic Journal* 93(4): 763–783.

Weitzman, M.L. 1984. *The share economy*. Cambridge, MA: Harvard University Press.

Weitzman, M.L. 1985a. Profit sharing as macroeconomic policy. *American Economic Review: Papers and Proceedings* 75(2): 41–45.

Weitzman, M.L. 1985b. The simple macroeconomics of profit sharing. *American Economic Review* 75(5): 937–953.

Weitzman, M.L. 1986. Reply to Nuti. (In Italian.) *Politica ed Economia* 4, April.

Coghlan, Timothy (1855–1926)

Colin G. Clark

Coghlan was the son of a poor family in Sydney, and obtained his education through scholarships. Like several other distinguished economists, his

original career was in engineering, in this case civil engineering, which led to his taking an interest in statistics. New South Wales had been established as a penal colony under military rule, and therefore right from the colony's beginning people and governments were accustomed to statistical enumeration much more thorough than in the rest of the world. Coghlan was appointed statistician by the colonial government, and produced over a long period of years a large output of excellent statistical papers.

Before Federation in 1901, there were six separate colonial governments in Australia, often pursuing different objectives. This was particularly the case in the highly controversial matter of free trade or protection. The two most populous colonies were Victoria and New South Wales. Victoria had attracted a large population through the abundance of alluvial gold in its riverbeds. This was soon exhausted; and deep mining fell far short of taking its place. The Victorian Government considered that the only way to provide employment to make up for the decline in mining was to establish manufacturing industries under high tariff protection. New South Wales (which admittedly had not had such a destabilizing gold rush) was determined to adhere to free trade, with the object of promoting agricultural and pastoral production. Some of Coghlan's writings were polemic, strongly stating the free trade case. In the *Wealth and Progress of New South Wales*, in succeeding editions, the available statistics of the colony were assembled, pointing to a superior rate of growth to that of Victoria.

Coghlan was fifty years ahead of his time in making national product estimations for New South Wales. Without any experience in any other country to guide him, he devised methods which would pass muster today. His work however was not followed up. There was no further estimate of Australian national product until the study by Benham in the 1920s, using less precise methods, and then another gap until Sir John Crawford and I prepared a study in 1937.

Coghlan prepared a series of publications on a more extensive basis, *The Seven Colonies of Australasia* (i.e. including New Zealand; it was thought at the time New Zealand might enter the Australian federation).

The free trade case was lost in 1908, when Victoria, with its allies in some other states, persuaded the new federal government to impose highly protective tariffs.

Coghlan was posted to London, where he performed most of the duties (but without the title, which went to the political head) of Agent General for New South Wales. The duties were varied and responsible, including the raising of large sums in loans.

In partial retirement, Coghlan embarked upon a *magnum opus*, *Labour and Industry in Australia*, which was published in England in 1918, giving a most full and detailed account of its subject, but illuminated by stories and anecdotes which make it readable.

Selected Works

1893. *Sheep and wool in New South Wales, with history and growth of the pastoral industry of the colony as regards both these items of production*. Sydney: Potter.

1898a. *Notes on the financial aspect of Australian Federation*, n.t.p. Sydney: W.A. Gullick.

1898b. *Notes on the financial aspect of Australian Federation. The Position of Tasmania and Western Australia*. Sydney: W.A. Gullick.

1898c. *Notes on the financial aspect of Australian Federation. The incidence of the federal tariff*. Sydney: W.A. Gullick.

1898d. *Tables to accompany notes on financial aspects of federation*. Sydney: W.A. Gullick.

1898e. *A statistical account of the seven colonies of Australasia 1861–1897*. Sydney: W.A. Gullick.

1902. *The progress of Australasia in the nineteenth century*. London, Philadelphia: The Linscott Publishing Company.

1903a. (With T.T. Ewing.) *The progress of Australia in the century*. London/Edinburgh: W. & R. Chambers Ltd.; Philadelphia/Detroit: The Bradley-Garretson Co., Ltd.

1903b. *The decline in the birth-rate of New South Wales and other phenomena of childbirth: An essay in statistics*. Sydney: W.A. Gullick.

1918. *Labour and industry in Australia, from the first settlement in 1788 to the establishment of the commonwealth in 1901*. London/New York: Oxford University Press.

Cognitive Ability

William T. Dickens

Abstract

Modern psychological theory views cognitive ability as multidimensional while acknowledging that the many different abilities are themselves positively correlated. This positive correlation across abilities has led most psychometricians to accept the reality of a general cognitive ability that is reflected in the full scale score on major tests of cognitive ability or IQ. This article provides an introduction to the history of cognitive testing and some of its major controversies. Evidence supporting the validity of measures of cognitive ability is presented and the nature and implications of group differences are discussed along with evidence on its malleability.

Keywords

Ability tests; Achievement tests; Cognitive ability; Cultural bias; External validity; Factor analysis; Heritability; IQ; Intelligence; Stereotype threat

JEL Classifications

D1

Some people are obviously and consistently quicker than others to understand new concepts; they solve unfamiliar problems faster, see relationships that others don't and are more knowledgeable about a wider range of topics. We call such people smart, bright, quick, or intelligent. Psychologists have developed tests to measure this trait. Originally called IQ tests (for Intelligence Quotient because the measures were constructed as the ratio of mental age to chronological age multiplied by 100), that name has fallen out of favour. Instead, such tests are now often referred to as tests of cognitive ability. Although the term IQ is still sometimes used to refer to what such tests measure, none constructs a ratio.

History

Spearman (1904) first popularized the observation that individuals who do well at one type of mental task also tend to do well at many others. For example, people who are good at recognizing patterns in sequences of abstract drawings are also good at quickly arranging pictures in order to tell a story, telling what three-dimensional shapes drawn in two dimensions will look like when rotated, tend to have large vocabularies and good reading comprehension, and are quick at arithmetic. This pattern of moderate to strong positive correlations across the whole spectrum of mental abilities led Spearman to hypothesize the existence of a general mental ability similar to the common notion of intelligence. A person's ability with any particular type of task would be equal to the sum of that person's general ability plus considerations unique to that particular task. Thus general ability could be measured by constructing sub-tests of a number of similar items (individual tasks of the same type such as arithmetic problems) of differing complexity. Each sub-test would present items of a different type, and individual scores across sub-tests could be aggregated. Task specific factors would average out leaving the final score as mainly a measure of general ability or 'g'. Using an approach like this Binet (1905) developed the first IQ test as a way of identifying students' academic potential. That test was adapted for use in English by Terman and in 1916 became the Stanford–Binet IQ tests – still one of the most commonly administered tests of cognitive ability.

Spearman's hypothesis of a single general mental ability and many specific abilities was challenged by Thurstone (1935), who popularized

the notion that people had a number of independent primary mental abilities rather than a single general mental ability. Both Spearman and Thurstone made contributions to the development of factor analysis as a way to identify the presence of unobserved variables (abilities) that affect a number of observable variables (sub-test or item scores). Today, the Spearman–Thurstone debate has been resolved with a compromise. The most common view among psychometricians who study cognitive ability is that there are a number of different abilities. Some people are better at solving problems verbally while others are good at solving problems that involve visualization. Some people who are good at both of these things may be only average at tasks that rely heavily on memory. However, there is a tendency for people who perform well in any of these broad areas to perform well in all others as well (Carroll 1993). Most modern tests of cognitive ability provide both a full-scale score that is most reflective of general intelligence, and a number of special-ability specific sub-scores as well.

Validity

Binet's is considered the first successful test of cognitive ability in that it was able to accurately predict teachers' assessments of their students on the basis of a relatively short verbally administered test. Scores on tests of cognitive ability correlate well with common perceptions of how bright or smart someone is. They are also strongly correlated with measures of academic achievement such as achievement test scores, grades and ultimate educational attainment (typically .5 or better). They are less highly correlated (.5 or less) with many important life outcomes including reported annual income and job status. Performance on a wide range of jobs and work tasks is positively related to cognitive test scores with performance on more demanding jobs having higher correlations. Some have claimed that general cognitive ability is responsible for most of this explanatory power (Ree and Earles 1992; Ree et al. 1994). This was a major theme of the controversial best-seller *The Bell Curve* (Herrnstein and Murray 1994). Heckman (1995), in a review of that book, argues that even though g has significant explanatory power, many other factors, both cognitive and non-cognitive, matter as well.

Finally, test scores are correlated with a number of social behaviours including unwed motherhood, criminal activity, and welfare receipt (Jensen 1998, ch. 9). While these correlations are substantial, and cognitive test scores are typically better predictors of most of these outcomes than any other single personal attribute, they still explain less than half the variance.

Individuals' scores on tests of cognitive ability also tend to be strongly correlated over time – much more so for adults than for children. A study of older adults found their full-scale IQ scores to be correlated .92 when tested at two points in time three years apart (Plomin et al. 1994). In contrast, a study of children tested at two points in time roughly two years apart found correlations of only .46 for those who were less than one year old at first testing and .76 for those who were one year old at first testing (Johnson and Bradley-Johnson 2002).

It is common to draw a distinction between tests of achievement and tests of ability. Achievement tests measure how much knowledge the test taker has accumulated in a particular area while ability tests endeavour to measure how quickly a person can solve unfamiliar problems. Typically, scores on the two types of tests are highly correlated. In fact, all tests of ability are, to some degree, tests of achievement as it is impossible to measure ability without also measuring the test taker's reading or verbal comprehension at least. Further, to the extent that the task being tested relies on knowledge of geometry, arithmetic, general knowledge, and so on, the rolls of the achievement test and ability test are confounded.

Cultural bias has been a concern with knowledge-based tests. Some knowledge is more accessible to some people than others. For example, we would expect that a child growing up with upper middle-class parents in New York or Paris to find it easier to learn the distance between the two cities (a general knowledge question that was once on one of the popular IQ tests) than someone from the slums of St. Louis or a tribesman from

the bush in Africa. For this reason a number of tests have been constructed that require a minimal amount of prior knowledge, such as Cattell's culture fair test (Cattell 1960) or Raven's progressive matrices (Raven 1941).

Group Differences

No matter what test is administered, men and women of the same background tend to have very similar average scores on tests of cognitive ability, though they differ slightly in their performance on some sub-tests (Jensen 1998, pp. 531–6). However, there are large differences across ethnic groups and geographic areas. The difference that has generated the most controversy is the difference in average scores of US blacks and whites, which is typically reported to be about one white standard deviation, though this gap has declined some in recent years (Dickens and Flynn 2006). Do these represent real differences in cognitive ability or do they reflect cultural bias in the tests?

Defenders of the tests offer several pieces of evidence suggesting that they are unbiased. Foremost is the evidence of 'external validity' – that the same regression equation that predicts outcomes such as job performance, grades, or educational attainment for one group will typically do a similarly good job for any other group. Also, different groups find the same questions more or less difficult. Members of different groups with similar scores will have similar patterns of right and wrong answers. If some questions are more culturally biased than others, the disadvantaged group should find those items more difficult than the mainstream group does. But researchers looking for such cultural bias have found no evidence of it (an exception occurs when one of the groups being compared is made up of non-native speakers of the language in which the test was administered, in which case scores on questions requiring a better knowledge of the language will be lower). Surprisingly, to the extent that there are black–white differences across test items, blacks do worse on what seem to be some of the least culturally dependent items – those involving abstract or symbolic problem-solving. Differences tend to be smaller on seemingly culturally rich items such as general knowledge. Herrnstein and Murray (1994, Appendix 5) provide a review of the evidence on bias.

The best evidence that tests can be biased in at least some circumstances comes from studies of a phenomenon called stereotype threat. It has been shown that reminding people of their group identity can cause them to perform in ways more consistent with stereotypes of the group's abilities. For example, blacks have been found to perform worse on some particularly difficult vocabulary items when given a questionnaire that asked them to state their race before taking the test or when the test was represented as a test of intelligence as opposed to a test of vocabulary. Women who were told that the difficult math test they were taking generally showed gender differences performed worse than those taking the same test who were told the test showed no differences. Men showed the opposite effect and performed better when told the test showed a gender difference (Steele 1997). However, it has not been demonstrated that stereotype threat produces substantial bias on standard tests in standard test-taking circumstances.

While most evidence is consistent with the view that tests provide a fair measure of the underlying concept of cognitive ability across ethnic groups, it is not conclusive. For example, since tests rarely explain as much as half the variance in the outcomes in studies of external validity, there is always the possibility that the tests underestimate black cognitive ability but that other disadvantages pull down black performance. If true, the validity of the tests as predictors of practical outcomes is an artifact of offsetting biases. This could explain why it is that when regressions of white performance on white test scores fail to predict black performance they tend to predict better performance than is observed. Further, commonsense notions that people from different cultural backgrounds probably have less opportunity to acquire certain types of information or practise certain skills should be given some weight. If studies find that blacks do no worse than similarly scoring whites on highly culturally loaded items, that could indicate that the poorscoring whites

were similarly disadvantaged. If disadvantage is more common for blacks than whites due to discrimination, that disadvantage could still explain some of the score gap. However, the strong correlation of even the culturally reduced tests with performance, and the similar magnitude of the gap on those tests between groups, suggest that much of the measured gap in ability between groups reflects real differences in average developed ability. This conclusion naturally leads to the consideration of the sources of those differences.

The question of whether individual, and particularly group, differences in cognitive ability are due more to nature or to nurture has been enormously controversial for the last century. Dickens (2005) presents a summary of the evidence on the origin of black–white differences and concludes that they are most likely not substantially genetic in origin. Rushton and Jensen (2005) reach the opposite conclusion. Whatever the right answer, whether the black–white gap has genetic origins is probably the wrong question. It seems that people are concerned with the issue mainly because they confuse having a genetic cause with immutability. While genes almost certainly play a large role in explaining individual differences in cognitive ability within ethnic groups raised in similar circumstances, it also seems that developed cognitive ability is highly malleable.

Malleability

A large amount of evidence has accumulated on the role of genes in explaining individual differences in cognitive ability. Several reviews of this literature conclude that differences in genetic endowment explain somewhere between 60 per cent and 80 per cent of the variance in cognitive ability in representative samples of the adult population in developed countries. The percentage for children is lower than for adults, with most estimates placing it around 40 per cent for six-year-olds (Plomin et al. 2001; Neisser et al. 1996). The figure is also estimated to be lower among disadvantaged populations (Turkheimer et al. 2003) though not consistently (Asbury et al. 2005). This figure is referred to as the heritability of cognitive ability. It is estimated by contrasting people with different degrees of relatedness raised in the same home or people with similar relatedness raised in different homes. For example, the correlation of the cognitive ability of identical twins raised in completely independent environments will be equal to the heritability of cognitive ability under the assumptions typically employed to make such estimates. While this evidence establishes that genes play a large role in determining individual differences, little is known about which genes are involved or how they influence cognitive ability (Plomin et al. 2001).

The high heritability of cognitive ability has led some to conclude that people's environments play little role in shaping their ability and that, therefore, individual differences are largely immutable and group differences must be largely due to differences in average genetic endowment. It has been argued that, if all of the observable differences in environment between people produce only 40 per cent or less of the variance in cognitive ability, then the large differences between blacks and whites could not result from the relatively small differences in environment between the average white and the average black. Thus differences in genetic endowment must play a substantial role. A formal version of this argument was first presented by Jensen (1973, pp. 135–9). A similar argument was made by Herrnstein and Murray (1994, pp. 298–9).

Yet despite the high heritability of cognitive ability, it does seem to be quite sensitive to environmental changes. In a review of the effects of early education programmes, Lazar and Darlington (1982, p. 44) noted that 'The conclusion that a well-run cognitively oriented early education program will increase the IQ sores of low-income children by the end of the programs is one of the least disputed results in educational evaluation'. The gains they surveyed were often quite large, though they also tended to decline substantially after children left the programmes. There is also evidence that being in a cognitively demanding environment can increase measured cognitive ability. Ceci (1991) surveys the evidence on the effects of school attendance on measured ability and finds it to be substantial.

Finally, the most profound changes in measured cognitive ability have taken place over time. James Flynn has documented huge gains in cognitive ability – as much as a standard deviation or more a generation – in more than 14 countries. Numerous other authors have found gains on other tests and in other countries (Flynn 1987, 1998, 2006). This phenomenon of large and pervasive gains has been dubbed 'the Flynn Effect'.

How is it that large gains are possible in the face of high heritability estimates? The chief flaw in the argument that high heritability implies a limited role for environment is that it misunderstands what heritability is measuring. It ignores the possibility that genetic and environmental influences might be correlated. In particular, it ignores the possibility that genetic influences on ability are largely the work of environmental advantages that come about due to modest physiological advantages.

Consider a sports analogy. Identical twins raised apart have a shared genetic endowment that tends to make them notably taller than their peers. As such they are both better basketball players. Even though they are raised apart, both are likely to spend more time playing basketball than other children their age. They are good at it and thus enjoy it more than other activities in which they do not naturally excel. Consequently they both get more practice at basketball than their peers, and that makes them better at the game. Being better players than their peers, they are more likely to be picked by coaches for high-school teams and more likely to receive yet more practice and more intensive coaching. If this leads to them playing in college they will both be enormously better players than the average person. A small physiological difference, which would make only a very modest difference in their performance on the court if they were untrained and inexperienced, has mushroomed into a huge difference in performance because it has been reinforced by the environmental influences of practice and coaching.

It is not hard to imagine the same thing happening with cognitive ability. Someone who is slightly quicker or has an emotional disposition amenable to thought and contemplation will be more likely to spend more time in intellectual pursuits. Such a person will likely receive positive reinforcement from teachers and be more likely to be tracked into more demanding classes and to develop friendships with other similarly disposed children. Such a child will have much more opportunity to practise intellectual work and receive more 'coaching' in intellectual pursuits. A small initial physiological difference could mushroom into a large difference in ability through a process whereby the advantage leads to a better environment which improves ability and gives access to even better environments.

If such reciprocal causation is at work in the development of cognitive ability, then small persistent exogenous differences in environment could produce large differences in cognitive ability. Dickens and Flynn (2001) lay out a formal model of such a process. If in a cross section of people in the same ethnic group most exogenous environmental differences are transient, then they will not accumulate through reciprocal causation and will not explain much variance across individuals. However, small persistent differences between groups or generations could cause large differences if they drive the engine of reciprocal causation. Similarly, preschool programmes which enrich children's cognitive environment can have large effects, but once the children are removed from the programme the process can work in reverse and unravel the gains. The exogenous decline in the quality of the environment from the removal of the programme's stimulation sets off a downward spiral of poorer performance leading the child into poorer environments, yet poorer performance and so on.

Conclusion

Modern psychology views cognitive ability as having a number of dimensions, all of which seem to be correlated with one another. Many interpret this correlation as reflecting an underlying general cognitive ability, or g, that is measured by the full-scale scores on the major tests of cognitive ability or IQ. General cognitive ability is an important predictor of a wide range of economic and life outcomes, with similar predictive validity across groups with different average levels of ability. Still, cognitive test scores typically explain

far less than half the variance in life outcomes, so cognitive ability is only one important factor among many that explain success.

Adult differences in cognitive ability within representative samples of ethnic groups raised in similar circumstances are subject to substantial genetic influence, but this does not mean that group differences are genetic in origin. Despite the large role played by genetic differences in explaining adult variance in cognitive ability, there is considerable evidence that intelligence is highly malleable and the life outcomes influenced by intelligence even more so.

See Also

► Behavioural Genetics

Bibliography

Asbury, K., T. Wachs, and R. Plomin. 2005. Environmental moderators of genetic influence on verbal and nonverbal abilities in early childhood. *Intelligence* 33: 643–661.

Binet, A. 1905. New methods for the diagnosis of the intellectual level of subnormals. *L'Année Psychologique* 12, 191–244. English edition: 1916. *The Development of Intelligence in Children* (trans: Elizabeth S. Kite). Vineland: Publications of the Training School at Vineland.

Carroll, J. 1993. *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge, MA: Cambridge University Press.

Cattell, R. 1960. *Measuring intelligence with culture fair tests*. Champaign: Institute for Personality and Ability Testing.

Ceci, S. 1991. How much does schooling influence general intelligence and its cognitive components? A reassessment of the evidence. *Developmental Psychology* 27: 703–722.

Dickens, W. 2005. Genetic differences and school readiness. *The Future of Children* 15: 55–69.

Dickens, W., and J. Flynn. 2001. Heritability estimates versus large environmental effects: The IQ paradox resolved. *Psychological Review* 108: 346–369.

Dickens, W., and J. Flynn. 2006. Black Americans reduce the racial IQ gap: Evidence from standardization samples. *Psychological Science* 17: 913–920.

Flynn, J. 1987. Massive gains in 14 nations: What IQ tests really measure. *Psychological Bulletin* 101: 171–191.

Flynn, J. 1998. IQ gains over time: toward finding the causes. In *The rising curve long-term gains in IQ and related measures*, ed. U. Neisser. Washington, DC: American Psychological Association.

Flynn, J. 2006. Efeito Flynn: repensando a inteligência e seus efeitos [The Flynn effect: rethinking intelligence and what affects it]. In *Introdução à Psicologia das Diferenças Individuais [Introduction to the psychology of individual differences]*, ed. C. Flores-Mendoza and R. Colom. Porto Alegre: ArtMed.

Heckman, J. 1995. Lessons from the bell curve. *Journal of Political Economy* 103: 1091–1120.

Herrnstein, R., and C. Murray. 1994. *The Bell Curve*. New York: Free Press.

Jensen, A. 1973. *Educability and group differences*. New York: Harper and Row.

Jensen, A. 1998. *The g factor: The science of Mental Ability*. Westport: Praeger.

Johnson, C., and S. Bradley-Johnson. 2002. Construct stability of the cognitive abilities scale-second edition for infants and toddlers. *Journal of Psychoeducational Assessment* 20: 144–151.

Lazar, I., and R. Darlington. 1982. Lasting effects of early education: A report from the consortium for longitudinal studies. In *Monographs of the Society for Research in Child Development*, vol. 47. Chicago: Chicago University Press.

Neisser, U., G. Boodoo, T. Bouchard Jr., A. Boykin, N. Brody, S. Ceci, D. Halpern, J. Loehlin, R. Perloff, R. Sternberg, and S. Urbina. 1996. Intelligence: Knowns and unknowns. *American Psychologist* 51: 77–101.

Plomin, R., J. Defries, G. McClearn, and P. McGuffin. 2001. *Behavioral genetics*. 4th edn. New York: Worth Publishers.

Plomin, R., N. Pedersen, P. Lichtenstein, and G. McClearn. 1994. Variability and stability in cognitive abilities are largely genetic later in life. *Behavior Genetics* 24: 207–215.

Raven, J. 1941. Standardization of progressive matrices. *British Journal of Medical Psychology* 19: 137–150.

Ree, M., and J. Earles. 1992. Intelligence is the best predictor of job performance. *Current Directions in Psychological Science* 1: 86–89.

Ree, M., J. Earles, and M. Teachout. 1994. Predicting job performance: Not much more than g. *Journal of Applied Psychology* 79: 518–524.

Rushton, J., and A. Jensen. 2005. Thirty years of research on race differences incognitive ability. *Psychology, Public Policy, and Law* 11: 235–294.

Spearman, C. 1904. General intelligence, objectively determined and measured. *American Journal of Psychology* 15: 201–293.

Steele, C. 1997. A threat in the air: how stereotypes shape intellectual identity and performance. *American Psychologist* 52: 613–629.

Thurstone, L. 1935. *The vectors of the mind*. Chicago: University of Chicago Press.

Turkheimer, E., A. Haley, M. Waldron, B. D'Onofrio, and I. Gottesman. 2003. Socioeconomic status modifies heritability of IQ in young children. *Psychological Science* 14: 623–628.

Cohen Stuart, Arnold Jacob (1855–1921)

Arnold Heertje

Keywords

Cohen Stuart, A. J.; Cournot, A. A.; Marginal utility of income; Oligopoly; Optimal taxation; Pierson, N. G.; Progressive and regressive taxation

JEL Classifications

B31

Born in The Hague, Cohen Stuart was an engineer who took up the challenge put forward by the famous Dutch economist and politician N.G. Pierson to study the mathematical foundations of what we would call nowadays an optimal tax structure. His thesis (Cohen Stuart 1889) has been reprinted in part (Musgrave and Peacock 1958).

The international attention to Cohen Stuart's exposition is due to the thorough discussion by F.Y. Edgeworth in his article on the pure theory of taxation (Edgeworth 1897). Following a lead by Pierson, Cohen Stuart studied the impact of the principle that each taxpayer should sacrifice an equal proportion of the total utility which he derives from material resources. He proved that it depends on the decrease of marginal utility of income, whether the income taxed above a certain minimum will be progressive, regressive or proportional in relation to the level of income. Cohen Stuart argues that in most practical cases a modest progressive tax rate will emerge.

Although based on old-fashioned concepts of measurable utility, Cohen Stuart's contribution to the analysis of the optimal income tax is part of the modern theory of optimal taxation (Mirrlees 1971) and therefore comparable to Cournot's role in the development of the theory of oligopoly.

See Also

▶ Pierson, Nicolaas Gerard (1839–1909)

Selected Works

1889. *Bijdrage tot de theorie der progressieve inkomstenbelasting*. The Hague: Nijhoff.

Bibliography

Edgeworth, F.Y. 1897. *Papers relating to political economy*. Vol. 2. London: Macmillan. 1925.

Mirrlees, J.A. 1971. An exploration in the theory of optimum income taxation. *Review of Economic Studies* 38: 175–208.

Musgrave, R.A., and A.T. Peacock, eds. 1958. *Classics in the theory of public finance*. London: Macmillan.

Cohen, Ruth Louisa (Born 1906)

Phyllis Deane

Ruth Cohen is one of the select band of leading professional economists whose importance is measured more by her emphatically commonsense influence on colleagues and pupils than by the length of her publications list. Born in 1906, she entered Newnham College, Cambridge, in 1926 to read for the Economics Tripos within a Faculty that contained a galaxy of outstanding individuals and included such stars as Keynes, Pigou, Robertson, Maurice Dobb and Piero Sraffa. She had already developed a research interest in agricultural economics when, after spending a couple of years as Commonwealth Fund Fellow at Stanford and Cornell, she joined the Oxford University Agricultural Economics Research Institute in 1933. For the next six years she pursued that specialization and completed two well-received books – one an analytical and statistical history of milk prices and the other a text on the economics of agriculture which J.M. Keynes invited her to write for the

Cambridge Economic Handbooks, of which he was then editor.

In common with most of the leading British economists of her generation she became a temporary civil servant soon after the outbreak of World War II, serving first at the Ministry of Food and later at the Board of Trade when questions of postwar reconstruction were rising to the top of the economic policy agenda of government. In 1945 she returned to Cambridge as lecturer in the Faculty of Economics and Politics and as Director of Studies in Economics at Newnham College, to which she had been elected a Fellow in 1939 and of which she was to become an active Principal in 1954. Soon after retiring from her university post in 1972 she was elected to the Cambridge City Council, where for a decade and a half she has devoted her formidable energy and talent as an applied economist to local government problems.

Ruth Cohen's reputation among contemporary economists has rested largely on her capacity to offer forceful, direct and perceptive oral comments on issues of current economic debate – theoretical as well as applied. This is the kind of salutary influence on the discipline that is rarely acknowledged in print, except in the occasional footnote. In the event, however, her typically terse intervention in the torrid capital theory debates that raged in the learned journals of the 1950s and 1960s has been duly credited with having triggered off a spate of articles in the capital-switching and capital-reversing phase of the debate – a phase which eventually ended in general agreement that her observation had revealed what some would describe as a fatal flaw, and others an awkward anomaly, in the orthodox neo-classical theory of the production function. To have stimulated this degree of consensus in a theoretical controversy which has carried unusually heavy methodological and ideological undertones is no mean achievement.

Selected Works

1936. *A history of milk prices*. Oxford: Agricultural Economics Research Institute.
1940. *The economics of agriculture*. Cambridge: Cambridge University Press.

References

Harcourt, G.C. 1972. *Some Cambridge controversies in the theory of capital*. Cambridge: Cambridge University Press.
Johnson, H.G. 1978. Ruth Cohen: A neglected contributor to contemporary capital theory. In *The shadow of Keynes*, ed. E.S. Johnson. Oxford: Blackwell.

Cohn, Gustav (1840–1919)

H. C. Recktenwald

Lecturer (*Privatdozent*) at the University of Heidelberg in 1869, Cohn was appointed professor of economics at the Riga Polytechnic Institute (1869–72). After spending some years in England he was appointed to a chair of economics at the Zürich Polytechnic Institute in 1875 and then in Göttingen in 1884. There he lived up to his death in 1919.

Cohn is noted for his pioneering contributions to the theory and policy of transportation and public finance. In his *Untersuchungen* (1874–75), *Eisenbahnpolitik* (1883) and *System* (1898, vol. 3), utilizing biased materials produced by parliamentary commissions, he strongly recommended railway centralization and government ownership while opposing canal construction; yet he failed to test the efficiency of these policy recommendations. In the field of public finance he is (like Rau, Roscher, von Stein and Wagner) the typical German exponent of Smith's liberal principles (not his moral theory) coupled with his own historical and ethical ideas, which were based on relatively poor analysis and synthesis. Cohn attributed to the state an economic and moral competence which he unquestioningly assumed. Advocating the legitimacy of value judgements (*Werturteile*) and ethical norms in economic science, he dealt with equity in taxation, particularly the controversial subsistence level and progressive taxation. Seligman rightly classes Cohn among the founders of the science of public finance.

His writings on general economics (1885–98, 1886) are distinguished by a philosophical foundation and a brilliant essayistic style which earned him a great reputation among his contemporaries.

Selected Works

1874–5. *Untersuchungen über die englische Eisenbahnpolitik*. Leipzig: Duncker & Humblot.
1883. *Die englische Eisenbahnpolitik der letzten zehn Jahre (1873–1883)*. Leipzig: Duncker & Humblot.
1885–98. *System der Nationalökonomie*. 3 vols, Stuttgart: Enke.
1886. *Nationalökonomische Studien*. Stuttgart: Enke.

Bibliography

Recktenwald, H.C. (ed.). 1973. *Political economy: A historical perspective*. London: Collier-Macmillan.

Cointegration

Mark W. Watson

Abstract

This article summarizes the mathematical structure of cointegrated time series models and discusses econometric procedures commonly used to analyse cointegrated time series. This discussion is carried out in the context of stochastic trends that follow driftless I(1) or 'unit root' processes. The article concludes with a brief discussion of cointegration in the context of more general stochastic trends.

Keywords

Beveridge–Nelson decomposition; Cointegration; Dickey–Fuller unit root tests; Error-correction terms; Granger, C.W.J.; Heteroskedasticity and autocorrelation corrections; Maximum likelihood; Trend/cycle decomposition; Unit roots; Variance; Vector autoregressions; Vector error correction model; Vector moving average models

JEL classifications

C32

Cointegration means that two or more time series share common stochastic trends. Thus, while each series exhibits smooth or trending behaviour, a linear combination of the series exhibits no trend. For example, short-term and long-term interest rates are highly serially correlated (so they are smooth and in this sense exhibit a stochastic trend), but the difference between long rates and short rates – the 'term spread' – is far less persistent and shows no evidence of a stochastic trend. Long rates and short rates are cointegrated.

The concept of cointegration was formalized by Clive W.J. Granger in a series of papers in the 1980s (Granger 1981; Granger and Weiss 1983; Granger 1986; Engle and Granger 1987), and in 2003 Granger received the Nobel Prize in Economics for this work. A flurry of research activity followed Granger's original contributions in this area and produced a practical set of econometric procedures for analysing cointegrated time series.

Mathematical Structure of I (1) Cointegrated Models

Let X_t denote a scalar I(1) stochastic process, with moving average representation $X_t = c(L)\varepsilon_t$, where ε_t is a scalar white noise process, and $c(L) = \sum_{i=0}^{\infty} c_i L^i$ is a polynomial in the lag operator L, and where the moving average coefficients, c_i decay sufficiently rapidly so that $\sum_{i=1}^{\infty} i|c_i| < \infty$. The Beveridge–Nelson decomposition (see trend/cycle decomposition) implies that X_t can be represented as $X_t = \tau_t + a_t$, where T_τ is a random walk, so that $\tau_t = \tau_{t-1} + e_t$, where e_t is white noise and a_t has a moving average representation $a_t = d(\mathrm{L})\varepsilon_t$, where

$\sum_{i=1}^{\infty} |d_i| < \infty$. Thus, X_t can be expressed as the sum of a stochastic trend, τ_t, and an I(0) process, a_t.

When X_t is an $n \times 1$ vector of I(1) processes, a similar result implies that $X_t = A\tau_t + a_t$, where A is a matrix of constants, τ_t is a vector of random-walk stochastic trends, and a_t is a vector of I (0) processes. Because X_t contains n elements, the vector τ_t will generally contain n stochastic trends. However, when τ_t contains only $k < n$ stochastic trends, A is $n \times k$, so that $\alpha' A = 0$, for any vector α in the null space of the column space of A. This means that $\alpha' X_t = \alpha' a_t$, so that the linear combination $\alpha' X_t$ does not depend on the stochastic trends. In this case, the time series making up X_t are said to be cointegrated. Any non-zero vector α that satisfies $\alpha' A = 0$ will annihilate the stochastic trend in $\alpha' X_t$, and vectors with this property are called cointegrating vectors. When A has full column rank, the number of linearly independent cointegrating vectors is $r = n - k$, which is called the cointegrating rank of the process.

For example, suppose that X_t contains $n = 3$ series representing interest rates on one-month, three-month and six-month US treasury bills. Suppose that $X_{it} = \tau_t + a_{it}$, for $i = 1, 2, 3$, where τ_t is a common stochastic trend shared by the three interest rates. Then $X_t = A\tau_t + a_t$, where $k = 1$ (there is a single stochastic trend), $A = (1\ 1\ 1)'$ (the trend has an equal effect on each of the interest rates) and $\alpha_1 = (1\ 0\ -1)'$ and $\alpha_2 = (1\ 0\ -1)'$ are two linearly independent cointegrating vectors, so that $r = 2$ and $\alpha_1' X_t$ and $\alpha_2' X_t$ denote the interest rate term spreads.

Vector moving average models (VMAs) and vector autoregressions (VARs) are often used to represent the linear properties of vector stochastic processes. The Granger representation theorem (see Engle and Granger 1987) shows that VMAs and VARs for cointegrated processes have special structures. In general, the VMA for an I(1) vector process is "$X_t = D(\mathrm{L})\varepsilon_t$, where ε_t is white noise with full rank covariance matrix. When X_t is not cointegrated, the $n \times n$ matrix $D(1)$, which contains the sum of the moving average coefficients, has rank n. But, when X_t is cointegrated, $D(1)$ has rank $k < n$, where k denotes the number of stochastic trends. When X_t is not cointegrated, the VAR for X_t can be written in terms of ΔX_t and has the form $\Phi(\mathrm{L})X_t = \varepsilon_t$, where $\Phi(\mathrm{L})$ is a stable lag polynomial (so its roots are outside the unit circle) and ε_t is white noise. When X_t is cointegrated, the VAR has the form Φ (L) $X_t = \beta\alpha' X_{t-1} + \varepsilon_t$, where α is an $n \times r$ matrix with columns that are the linearly independent cointegrating vectors. Thus, the cointegrated VAR expresses the elements of ΔX_t as functions of its own lags, but also includes the r regressors $\alpha' X_{t-1}$ in each of the VAR's n equations. The variables $a' X_{t-1}$ are called 'error-correction terms' and the cointegrated VAR is called a 'vector error correction model' (VECM).

Watson (1994) provides a summary of the algebra linking these various representations of the cointegrated model.

Testing for Cointegration

The time series making up X_t are cointegrated if the linear combinations $\alpha' X_t$ are I(0) random variables. If X_t is not cointegrated, then $\alpha' X_t$ will be I (1) for any non-zero vector α. Tests of cointegration ask whether $\alpha' X_t$ is I(1) or I(0).

Consider the simple case in which there is only one potential cointegrating vector, so that $\alpha' X_t$ is a scalar. Cointegration can then be tested using a unit root test applied to $\alpha' X_t$. The straightforward application of a unit root test requires that α is known, so that the scalar variable $\alpha' X_t$ can be calculated directly from the data. This is possible in many empirical applications (such as the interest rate example described above) where the value of α can be pre-specified.

Thus, suppose that α is known, and consider the competing hypotheses $\mathrm{H}_{\mathrm{I}(1)}$: $\alpha' X_t$ is I(1) and $\mathrm{H}_{\mathrm{I}(0)}$: $\alpha' X_t$ is I(0). The hypothesis $\mathrm{H}_{\mathrm{I}(1)}$ means that the elements of X_t are not cointegrated and the hypothesis $\mathrm{H}_{\mathrm{I}(0)}$ means that the elements are cointegrated. Under $\mathrm{H}_{\mathrm{I}(1)}$ the autoregressive model for $\alpha' X_t$ contains a unit root, while under $\mathrm{H}_{\mathrm{I}(0)}$, the autoregressive model for $\alpha' X_t$ is stable.

The null $\mathrm{H}_{\mathrm{I}(1)}$ can be tested against the alternative $\mathrm{H}_{\mathrm{I}(0)}$ using an augmented Dickey–Fuller

(ADF) unit root test or the modified ADF test developed in Elliott, Rothenberg and Stock (1996). The null $H_{I(0)}$ can be tested against $H_{I(1)}$ using the best local test proposed by Nyblom (1989), modified for serial correlation as described in Kwiatkowski et al. (1992), or a point-optimal test as discussed in Jansson (2004). (There are important practical considerations associated with the choice of the long-run-variance estimator (see heteroskedasticity and autocorrelation corrections) used in tests for the $H_{I(0)}$ null hypothesis because of the high degree of serial correlation under the alternative. See Müller (2005) for discussion.)

When α is not known, the unit root tests described in the last paragraph use $\hat{\alpha}'X_t$ in place of $\alpha'X_t$, where $\hat{\alpha}$ is an estimator of α. For example, Engle and Granger (1987) suggest estimating α by regressing the first element of X_t onto the other elements of X_t using OLS, and carrying out an ADF test using the residuals from this regression. Estimation of α changes the distribution of the ADF test statistic from what it is when α is known, so that critical values for the Engle–Granger test are different than the standard ADF critical values. As described in Phillips and Ouliaris (1990) and Hansen (1992) the correct critical values depend on the number of elements in X and on the properties of the deterministic trends in the model. Stock and Watson (2007) tabulate choices of critical values from the Phillips and Ouliaris (1990) and Hansen (1992) papers that are appropriate for data that follow I(1) processes that may or may not contain drift, and thus serve as conservative critical values. Modifications for tests of the $H_{I(0)}$ null versus the $H_{I(1)}$ alternative are discussed in Shin (1994) and Jansson (2005).

The tests outlined above are useful for testing whether a single series $\alpha'X_t$ is I (0) or I(1), but in many applications there may be more than one potential cointegrating relation ($r > 1$) so that it is useful to have tests for hypothesis that postulate different values of r. That is, it is useful to entertain hypotheses of the form $H_j : r = j$, for $j = 0, 1, \ldots n$. The hypothesis $r = 0$ means that there is no cointegration, $r = 1$ means that there is a single cointegrating vector, and so forth. As discussed in Johansen (1988), these tests are easily formulated and carried out using the VECM model. Recall that the VECM model has the form $\Phi(L)X_t = \beta\alpha' X_{t-1} + \varepsilon_t$. Consider the null and alternative hypotheses $H_o : r = r_o$ *vs.* $H_a : r = r_a$ where $r_a > r_o$, and write the VECM as $\Phi(L)X_t = \beta_o\alpha_0' X_{t-1} + \tilde{\beta}\tilde{a}'X_{t-1} + \varepsilon_t$, where α_o contains the r_o cointegrating vectors under the null and $\tilde{a}$ contains the additional cointegrating vectors under the alternative. Under the null hypothesis, the variables $\tilde{a}'X_{t-1}$ do not enter the VECM, while under the alternative these variables enter the VECM. Thus, the null and alternative can be written as $H_o : \tilde{\beta} = 0$ versus $H_a : \tilde{\beta} \neq 0$. As in the case of $r = 1$, the tests depend on whether the cointegrating vectors are known or unknown. When the cointegrating vectors are known, the regressors $\alpha_o'X_{t-1}$ and $\tilde{\alpha}'X_{t-1}$ can be constructed from the data, and the Wald test for $\tilde{\beta} = 0$ can be constructed using the usual regression formula. When the cointegrating vectors are unknown, the testing problem is more difficult, but Johansen (1988) provides an simple formula for the likelihood ratio test statistic. In either case, the critical values for the test are 'non-standard', that is, they are not based on the χ^2 or F distributions. Critical values for the tests depend on the values of $r_a - r_o$, the number of cointegrating vectors that are known and unknown, and the presence or absence of constants and time trends in the model. The various critical values are tabulated in Horvath and Watson (1995).

Estimating Unknown Cointegrating Coefficients

Unknown coefficients in cointegrating vectors are typically estimated using least squares and Gaussian maximum likelihood estimators (MLEs). The properties of these estimators can be understood by considering a simple bivariate model

$$X_{1t} = \theta X_{2t} + \eta_{1t}$$
$$X_{2t} = X_{2t-1} + \eta_{2t}$$

where $\eta_t = [\eta_{1t}\eta_{2t}]' \sim$ iid N $(0, \Sigma)$. In this example, there is one common trend that coincides with X_{2t}, the cointegrating vector is $\alpha = (1 - \theta)'$ where θ is an unknown parameter, the error correction term is $\alpha' X_t = \eta_{1t}$ which is potentially correlated with the innovation in the common trend, η_{2t}, and the assumption of normality is used to motivate the Gaussian MLE of θ.

The OLS estimator of θ has several interesting properties (Stock 1987). Even though X_{2t} and η_{1t} are correlated, the OLS estimator is consistent; indeed it is 'super-consistent' in the sense that $\hat{\theta}^{OLS} - \theta \sim O_p\left(T^{-1}\right)\hat{\theta}^{OLS}$, so that $\hat{\theta}^{OLS}$ converges to θ faster than the usual $\sqrt{T}$ rate familiar from regressions involving I(0) variables. These results follow because, in the cointegrated model, the regressor X_{2t} is I(1) and therefore is much more variable than an I(0) regressor $\left(\sum_{t=1}^{T} X_{2t}^2 \sim O_p\left(T^{-2}\right)\right.$ in this I(1) regression instead of $O_p(T^{-1})$ in the usual I(0) regression), and the correlation between X_{2t} and η_{1t} is non-zero, but vanishes as the sample size becomes large. (The covariance is constant, but the variance of X_{2t} increases linearly with t, so the correlation vanishes as t increases.)

Despite these intriguing and powerful features, the OLS estimator has two properties that make it unsatisfactory for many uses. First, while OLS is consistent, the correlation between the regressor and error term induces a bias in the large sample distribution of the estimator, and this bias can be severe in sample sizes typically encountered in applied work (Stock 1987). Second, the large-sample distribution of the OLS estimator is non-normal, and this complicates statistical inference. For example, the standard interval $\hat{\theta}^{OLS} \pm 1.96SE\left(\hat{\theta}^{OLS}\right)$ does not provide a 95 per cent confidence set even in large samples. Interestingly, Gaussian maximum likelihood estimators share the super consistency properties of OLS, but do not suffer from these unsatisfactory properties (Johansen 1988; Phillips 1991).

To construct the Gaussian MLE, factor the joint density of $\{X_t\}_{t=1}^{T}$ into the density of $\left\{X_{1t}|\,(X_{2t})_{t=1}^{T}\right\}_{t=1}^{T}$ and the density of $\{X_{2t}\}_{t=1}^{T}$. The density of $\{X_{2t}\}_{t=1}^{T}$ does not depend on θ, and the density of $X_{1t}|(X_{2t})_{t=1}^{T}$ is characterized by the Gaussian linear regression $X_{1t} = \theta X_{2t} + \beta X_{2t} + v_t$, where β is the regression coefficient from the regression of η_{1t} onto η_{2t} $(=\Delta X_{2t})$, v_t is the error in this regression, and $v_t|(X_{2t})_{t=1}^{T} \sim$ iid $\mathrm{N}(0, \sigma_2)$. Simple calculations (Phillips 1991) can then be used to show that $\hat{\theta}^{MLE} - \theta \sim O_p\left(T^{-1}\right)$ and that $\hat{\theta}^{MLE}|(X_{2t})_{t=1}^{T} \sim \mathrm{N}(\theta, V)$, where V depends on $(X_{2t})_{t=1}^{T}$. Thus, $\hat{\theta}^{MLE}$ is consistent, is conditionally normally distributed and unbiased, and $\left(\hat{\theta}^{MLE} - \theta\right)/V^{1/2} \sim \mathrm{N}(0,\ 1)$, so that inference about θ can be carried out using standard methods associated with the Gaussian linear regression model. Thus, for example, $\hat{\theta}^{MLE} \pm 1.96\ SE\left(\hat{\theta}^{MLE}\right)$ provides a valid 95 per cent confidence set for θ, where $SE\left(\hat{\theta}^{MLE}\right)$ is computed using the usual regression formula.

While these results may appear quite special (X_t is bivariate and η_t is normally distributed and serially uncorrelated) they carry over to more general models with minor modifications. For example, X_{1t} and X_{2t} may each be vectors and the regression $X_{1t} = \theta X_{2t} + \beta X_{2t} + v_t$ becomes a multivariate regression. Under weak assumptions on the distribution of η_t, there is sufficient averaging so that $V^{-1/2}\left(\hat{\theta}^{MLE} - \theta\right) \to N(0, I)$, meaning that the assumption of normality for η is not critical (although $\hat{\theta}^{MLE}$ still refers to the MLE computed by maximizing the Gaussian likelihood). Serial correlation in η_t can be handled in a variety of ways. For example, Saikkonen (1991) and Stock and Watson (1993) consider the 'dynamic OLS' (DOLS) regression $X_{1t} = \theta X_{2t} + \Sigma_{i=-k}^{k}\ \beta_i X_{2t-i} + v_t$, which includes enough leads and lags of ΔX_{2t} to insure that v_t is (linearly) independent of $(X_{2t})_{t=1}^{T}$. Phillips and Hansen (1990) and Park (1992) develop adjustments based on long-run covariance matrix estimators, and Johansen (1988) derives the exact Gaussian MLE based on the VECM. Under general assumptions, all of the estimators are asymptotically equivalent.

Alternative Models for the Common Trends

The concept of cointegration involves variables that share common persistent 'trend' components. The statistical analysis outlined above utilized a particular model of the trend component, namely, the driftless unit root process $\tau_t = \tau_{t-1} + e_t$. Analysis of this model highlights many of the key features of cointegrated processes, but more general models are often needed for empirical analysis. For example, constant terms are often added to the model to capture non-zero means of error correction terms or drifts in the trend process. These constant terms change the distribution of test statistics for cointegration in ways familiar from the effect of constants and time trends in Dickey–Fuller unit root tests (see Hamilton 1994). Hansen (1992) and Johansen (1994) contain useful discussion of the key issues. Higher-order integrated processes (for example, I(2) processes) are discussed in Johansen (1995), Granger and Lee (1990), and Stock and Watson (1993). Hylleberg et al. (1990) discuss cointegration at seasonal frequencies. Robinson and Hualde (2003) and the references cited therein discuss cointegration in fractionally integrated models.

Elliott (1998) discusses cointegrated models in which the trend follows a 'nearunit-root' process – an AR process with largest autoregressive root very close to 1.0. (Formally, the asymptotics use a local-to-unity nesting with largest root AR root equal to $1-c/T$, where c is a constant.) Elliott shows that, while the basic cointegrated model remains unchanged in this case, the properties of Gaussian maximum likelihood estimators of unknown cointegrating coefficients change in important ways. In particular, the Gaussian MLEs are no longer conditionally unbiased, and confidence intervals constructed using Gaussian approximations $\left(\text{for example, } \hat{\theta}^{MLE} \pm 1.96SE\left(\hat{\theta}^{MLE}\right)\right)$ can be very misleading. Elliott's critique is important because small deviations from exact unit roots cannot be detected with high probability, and yet small deviations may undermine the validity of statistical inferences constructed using large-sample normal approximation applied to Gaussian MLEs.

Several papers have sought to address the Elliott critique by developing methods with good performance for a range of autoregressive roots close to, but not exactly equal to 1.0. For example, Wright (2000) argues that if θ_0 is the true value of a cointegrating coefficient, then $X_{1t} - \theta_0 X_{2t}$ will be I(0), but if θ_0 is not the true value then $X_{1t} - \theta_0 X_{2t}$ will be highly persistent. He suggests testing that $\theta = \theta_0$ by testing the $H_{I(0)}$ null for the series $X_{1t} - \theta_0 X_{2t}$. Alternative testing procedures in this context are proposed in Stock and Watson (1996) and Jansson and Moreira (2006).

See Also

- ▶ Heteroskedasticity and Autocorrelation Corrections
- ▶ Trend/Cycle Decomposition
- ▶ Unit Roots

Bibliography

Elliott, G. 1998. The robustness of cointegration methods when regressors almost have unit roots. *Econometrica* 66: 149–158.

Elliott, G., T.J. Rothenberg, and J.H. Stock. 1996. Efficient tests for an autoregressive unit root. *Econometrica* 64: 813–836.

Engle, R.F., and C.W.J. Granger. 1987. Co-integration and error correction: Representation, estimation, and testing. *Econometrica* 55: 251–276.

Granger, C.W.J. 1981. Some properties of time series data and their use in econometric specification. *Journal of Econometrics* 16: 121–130.

Granger, C.W.J. 1986. Developments in the study of co-integrated economic variables. *Oxford Bulletin of Economics and Statistics* 48: 213–228.

Granger, C.W.J., and T.H. Lee. 1990. Multicointegration. *Advances in Econometrics* 8: 71–84.

Granger, C.W.J., and A.A. Weiss. 1983. Time series analysis of error-correction models. In *Studies in econometrics, time series, and multivariate statistics, in honor of T.W. Anderson*, ed. S. Karlin, T. Amemiya, and L.A. Goodman. San Diego: Academic.

Hamilton, J.D. 1994. *Time series analysis*. Princeton: Princeton University Press.

Hansen, B. 1992. Efficient estimation and testing of cointegrating vectors in the presence of deterministic trends. *Journal of Econometrics* 53: 86–121.

Horvath, M.T.K., and M.W. Watson. 1995. Testing for cointegration when some of the cointegrating vectors are prespecified. *Econometric Theory* 11: 952–984.
Hylleberg, S., R.F. Engle, C.W.J. Granger, and B.S. Yoo. 1990. Seasonal integration and cointegration. *Journal of Econometrics* 44: 215–238.
Jansson, M. 2004. Stationarity testing with covariates. *Econometric Theory* 20: 56–94.
Jansson, M. 2005. Point optimal tests of the null of hypothesis of cointegration. *Journal of Econometrics* 124: 187–201.
Jansson, M., and M. Moreira. 2006. Optimal inference in regression models with integrated regressors. *Econometrica* 74: 681–714.
Johansen, S. 1988. Statistical analysis of cointegrating vectors. *Journal of Economic Dynamics and Control* 12: 231–254.
Johansen, S. 1994. The role of the constant and linear terms in cointegration analysis of non-stationary variables. *Econometric Reviews* 13: 205–229.
Johansen, S. 1995. A statistical analysis of cointegration for I(2) variables. *Econometric Theory* 11: 25–59.
Kwiatkowski, D., P.C.B. Phillips, P. Schmidt, and Y. Shin. 1992. Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics* 54: 159–178.
Müller, U.K. 2005. Size and power of tests for stationarity in highly autocorrelated time series. *Journal of Econometrics* 128: 195–213.
Nyblom, J. 1989. Testing for the constancy of parameters over time. *Journal of the American Statistical Association* 84: 223–230.
Park, J.Y. 1992. Canonical cointergrating regressions. *Econometrica* 60: 119–143.
Phillips, P.C.B. 1991. Optimal inference in cointegrated systems. *Econometrica* 59: 283–306.
Phillips, P.C.B., and B.E. Hansen. 1990. Statistical inference on instrumental variables regression with I(1) processes. *Review of Economic Studies* 57: 99–124.
Phillips, P.C.B., and S. Ouliaris. 1990. Asymptotic properties of residual based test for cointegration. *Econometrica* 58: 165–193.
Robinson, P.M., and J. Hualde. 2003. Cointegration in fractional systems of unknown orders. *Econometrica* 71: 1727–1766.
Saikkonen, P. 1991. Asymptotically efficient estimation of cointegrating regressions. *Econometric Theory* 7: 1–21.
Shin, Y. 1994. A residual-based test of the null of cointegration against the alternative of no cointegration. *Econometric Theory* 10: 91–115.
Stock, J.H. 1987. Asymptotic properties of least squares estimates of cointegrating vectors. *Econometrica* 55: 1035–1056.
Stock, J.H., and M.W. Watson. 1993. A simple estimator of cointegrated vectors in higher-order integrated systems. *Econometrica* 61: 783–820.
Stock, J.H., and M.W. Watson. 1996. *Confidence sets in regression with highly serially correlated regressors*. Manuscript, Department of Economics, Princeton University.
Stock, J.H., and M.W. Watson. 2007. *Introduction to econometrics*. 2nd ed. Boston: Pearson-Addison Wesley.
Watson, M.W. 1994. Vector autoregression and cointegration. In *Handbook of economics*, ed. R.F. Engle and D.L. McFadden, vol. 4. Amsterdam: North-Holland.
Wright, J.H. 2000. Confidence sets for cointegrating coefficients based on stationarity tests. *Journal of Business and Economic Statistics* 18: 211–222.

Colbert, Jean-Baptiste (1619–1683)

D. C. Coleman

Keywords
Colbert, J.-B.; Colbertism; Mercantile system; Mercantilism

JEL Classifications
B31

Colbert was born at Reims on 29 August 1619 and died on 6 September 1683. In no way at all could he be called an economist. He was, however, one of the most powerful administrators, known to history, of measures affecting the economic life of a nation, to such an extent and with such lasting influence that his name is preserved in the notion of Colbertism.

He came of a mercantile family which had acquired some public offices. He learned his job as economic administrator by entering the service, in 1651, of a man he was effectively to succeed, Cardinal Mazarin. Once successfully installed in the service of Louis XIV, after Mazarin's death in 1661 his climb to power was rapid. He soon came to hold numerous offices of state: finance, commerce, buildings, the navy, and more besides. His achievements rested in part upon his exercising virtually undisputed power for 22 years as the dominant minister of the grandest of absolute

monarchs, and in part upon his own qualities of character which he brought to bear upon the economic problems of France as he perceived them. Those qualities included energy, tenacity, shrewdness, honesty, a notable ability to deploy the techniques of the courtier, and a wholly remarkable capacity for hard work. His hand was felt in every aspect of French economic life; and everywhere he exercised that passion for order which is so often the hallmark of the bureaucrat. Adam Smith sniffed at him as a 'laborious and plodding man of business ... accustomed to regulate the different departments of public offices' (Smith 1776, p. 627). But he was a lot more than that. Cold, humourless, and devoted, he was the super-servant of a super-king.

Those qualities did not, on the other hand, include any original economic ideas whatever. He had absorbed, with characteristic thoroughness, all the assumptions, maxims, dogmas, and assorted notions about economic matters which circulated in 16th- and 17th-century Europe, and to which the label of mercantilism has become attached. Consequently, by dint of his position and activities, and because a very large volume of his papers have survived for the historian, he has come down to posterity as the embodiment of conventional mercantilism in practice. Non-existent as a theoretical entity, mercantilism has acquired the appearance of a coherent economic policy probably more from Colbert's activities than from any other single historical source. And because it appeared, and was continued after his death, in the grandeur which was France, it was copied or adapted in other aspiring monarchies. French mercantilism or Colbertism thus became a recognizable reality in a way that the English 'mercantile system' did not.

The nature of his economic ideas can often be gathered from the explanatory memoranda which he addressed to Louis XIV (who was not always as interested in such matters as Colbert thought he should be). They have a familiar ring. He wanted money circulating in the kingdom, not because he identified money with wealth, but because it facilitated the payment of taxes and helped to stimulate economic activity; those branches of overseas trade which brought in precious metals were therefore to be especially favoured. Manufacturing industry deserved encouragement because it lessened French dependence on imports, because it was the basis of an export trade which brought in wealth, and because it employed the idle (the Catholic Colbert had the zeal for work and the disapproval of idleness normally thought of as peculiar to Puritanism). In the interest of the economic unification of France, internal trade and transport needed improvement by the removal of tolls and the repair of roads and bridges. Royal support was needed, and was secured, for the construction of canals – of which the most spectacular achievement was the opening in 1681 of the Canal des Deux Mers, providing a waterway between the Atlantic and the Mediterranean.

Colbert shared the pervasive belief in a fixed cake of trade, so that, as he patiently explained to Louis in March 1669, the whole trade of Europe was carried in a fixed number of vessels and therefore 'le commerce cause un combat perpétuel en paix et en guerre entre les nations de l'Europe, à qui on emportera la meilleure partie'. The Dutch, the English and the French were the 'acteurs de ce combat' (*Lettres* VI, p. 266). France's gain was to be secured by Holland's and/or England's loss. It followed that shipbuilding should be encouraged and the French navy and mercantile marine greatly enlarged. France should move in on trades hitherto dominated by her rivals. Hence his setting up in the 1660s of privileged trading companies: a French East India Company, a French West India Company to improve and exploit French colonies, and the Company of the North to tap the Baltic trade. Such views also provided an economic justification for the war which Louis launched against Holland in 1672. Colbert had to find the revenue for these and others of his master's military activities. Consequently, he devoted much time to trying to reform the royal finances. Many of his measures – for example, to improve the collection of taxes or to unify the customs system – were thus again part of a policy designed to improve the performance of the economy so that it could in turn yield more wealth to the greater glory of *le roi soleil.*

How much success attended Colbert's policies has been a matter of debate. Laissez-faire economists and economic historians of similar views have inevitably disparaged them and stressed the rigidities which were built into the French economy in the 18th century. His efforts to unify the chaotic diversity of French fiscal and customs administration were only very partially successful; his overseas trading companies were inadequately financed and generally unprofitable; his comparative neglect of agriculture left the basis of the economy in a poor state. But his work did greatly improve the size and efficiency of the French navy and mercantile marine; stimulate – albeit at a high cost – certain areas of French manufacturing industry; and encourage French merchant enterprise in branches of trade hitherto the preserve of others. Not all of this was evident in his own lifetime. But one thing was: Colbert died a very rich man, ennobled as Marquis de Seignelay, his brothers and sisters and cousins amply provided with lucrative sinecures, his sons as ministers or army officers, and his three daughters married off to dukes. Such were the 17th-century rewards of administering an economy.

Bibliography

Clément, P. (ed.). 1861–2. *Lettres, Instructions et Mémoires de Colbert*, 8 vols. Paris.

Cole, C.W. 1939. *Colbert and a century of French mercantilism,* 2 vols. New York: Columbia University Press.

Smith, A. 1776. In *An inquiry into the nature and causes of the wealth of nations*, ed. E. Cannan. New York: Modern Library. 1937.

Colbertism

D. C. Coleman

Colbertism is a term used to describe the economic policies associated with the French statesman, Jean-Baptiste Colbert; and sometimes, confusingly, as a synonym for mercantilist policies in general.

In the course of his account, and denunciation, of the mercantile system, Adam Smith presented it as something foisted upon governments by conspiring businessmen. Extending this view from England to France, he said of Colbert that he had been 'imposed upon by the sophistry of merchants and manufacturers' (Smith 1776, p. 434). Whatever degree of truth there may be in his account so far as it related to England – and there is some – it wholly misrepresents the mind of Colbert and the nature of Colbertism. Distrusting the self-interest of businessmen as a power for the greater good of society, Colbert believed profoundly that, although their pursuit of profits should be encouraged, the way to ensure that such activities redounded to the greater wealth, and hence power and glory, of France was by regulation and order. So Colbertism was essentially a systematic treatment of economic activities imposed from above by the King through his servant. It could be described as a version of the mercantile system appropriate to an absolutist state. It owed little or nothing to mercantile or manufacturing pressures brought to bear on governments. Although there were some similarities between Colbertism and English mercantilism, both in the ideas which lay behind it and in its outward forms as it affected overseas trade, the creation of Colbertian policies did not in the least resemble the process of bargaining and compromise between Crown and Parliament by which English mercantilism was muddled into existence. For this reason alone the term 'Parliamentary Colbertism', coined by Cunningham and used by him to describe English economic policy, 1689–1776 (Cunningham 1907, II, pp. 403–68), was singularly inappropriate. It was also inapt for the different reason that Colbertism was distinguished by a concern for the direct control of production which was wholly absent from the English version of mercantilist policies.

The quintessence of Colbertism is strikingly illustrated in Colbert's approach to manufactures. Observing that France had great industrial potential, with many and scattered crafts and substantial manpower, he set about the country's industrial

rehabilitation. He used a variety of weapons: subsidies, special tax reductions or exemptions, protection against foreign imports, the encouragement of early marriage and large families, grants of special privileges, and the establishment of *manufactures royales*. Disapproving, for example, of the way in which his countrymen imported and wore the woollen cloth or serges of Holland and England, he set up *manufactures royales* to stimulate their production in France; and in 1667 very sharply increased import duties against the offending English and Dutch imports. Similar techniques were used to promote the making of lace, silk stockings, tapestries, carpets, glassware, tinplate, soap, naval supplies, and cannon. Luxury items and textiles received particular attention. It has been said that 'the greatest industry in France was supplying the wants of the King and his court' (Cole 1939, II, p. 303). In quantitative terms this was probably untrue but its significance was very real; and such a statement could not possibly be made about English industry. Stimulation demanded regulation. So Colbert established a Code of Commerce, promulgated for textiles elaborate controls covering precise lengths, widths and other details of all types of textiles; established an apparatus of industrial inspection; and insisted upon all labour being organized within the guild structure.

Three points need to be stressed about these measures. First, Colbertism was here a continuation and codification, a new ordering of old practices; it was part of an *étatisme* with medieval roots. Second, at the time that Colbert was imposing these measures on the French economy, their English counterparts were withering away; the last legislative attempt at general regulation of the English cloth industry failed in 1678. Third, Colbert's regulative achievements were continued after his death: Colbertism brought many more detailed regulations in the seventy years after 1683.

Colbert's founding of privileged monopolistic trading companies shows a certain resemblance to the prior establishment of their counterparts in Holland and England. Again, however, the special nature of Colbertian mercantilism is evident both in the preponderance of royal and government finance in the early years of these companies because of inadequate mercantile enthusiasm for them; and in the degree of personal control which Colbert himself exercised, especially over the French East India Company. So far from being a product of mercantile pressures Colbertism ran foul of merchants on more than one occasion. Colbert made himself very unpopular with those of Marseilles, for example, when, obsessed by the need to keep money circulating so that taxes could be paid, he tried to prevent them from exporting coin in order to conduct their trade with the Levant. And the highly protective anti-Dutch tariff of 1667 attracted internal opposition because it so obviously invited retaliation.

The vast regulative apparatus built up by Colbert and his successors showed more contempt than understanding of the role of businessmen. French commercial and industrial advance during the 18th century, though owing something to Colbert's initiating stimuli, continued despite, rather than because of, the perpetuation of Colbertism. Indeed, one of the reasons for the final reaction against it was the extent to which the bureaucratic machine had become both corrupt in its operation and irrelevant to the needs of the French economy. It helped the proliferation in 18th-century France of a congerie of fiscal officeholders and a concomitant trade in offices and privileges functioning in and around an overblown court. Such practices certainly existed before Colbert's day; but just as Colbert brought a new administrative zeal to old economic ideas, so Colbertism came to provide a still more fertile soil for the growth of ancient corruptions. Meanwhile, however, it appealed to other states – Prussia and the German principalities, Russia, Austria, Spain – intent on building up or repairing economic bases for the support of absolutist courts, territorial ambitions, or the urge for military glory. The sorts of mercantilism which they adopted all varied a good deal, despite the common name and some common economic ideas. But those of central, eastern and southern Europe were often much nearer in spirit to Colbertism than to the mercantile system which Smith discerned in England or to the particular variety which the Dutch had erected in Holland. Colbertism was in this sense *sui generis*.

See Also

► Mercantilism

Bibliography

Cole, C.W. 1939. *Colbert and a century of French mercantilism*, 2 vols. New York: Columbia University Press.

———. 1943. *French mercantilism, 1683–1700*. New York: Columbia University Press.

Cunningham, W. 1907. *The growth of English industry and commerce,* 3 vols. Cambridge: Cambridge University Press.

Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations*. Ed. E. Cannan. New York: Modern Library edn, 1937.

Cole, George Douglas Howard (1889–1959)

Anthony Wright

A British socialist intellectual, G.D.H. Cole was born in Cambridge in 1889. He grew up in London and was educated at Balliol College, Oxford. As a young Oxford don, Cole came to prominence during the second decade of the century as a leading advocate of guild socialism (a doctrine of workers' control in industry) and adviser to the labour and trade union movements. After the collapse of guild socialism, Cole continued to be the outstanding socialist theorist and Labour Party intellectual in Britain during the interwar and immediate postwar periods, always combining academic work with political commitment. An encyclopaedist and polymath, Cole's published output was prodigious in both volume and range. He produced over a hundred books, and at different periods held academic posts in three disciplines (philosophy, economics, political theory) and could easily have held posts in at least two others.

Cole's central and lifelong preoccupation was with the advocacy of a decentralized, self-managing and participatory form of socialism. It is against this background that his work in economics has to be seen. Although he immersed himself in economic matters during the interwar period (when he was Reader in Economics at Oxford), he regarded this as a labour of necessity. In a basic sense, he did not *like* economics, and railed against the 'algebraic sterilities' of those economic theorists who divorced the subject both from social values and from the solution of pressing problems in the real world. He was, anyway, not equipped to enter the higher reaches of theoretical economics, and his own economic theory therefore remained essentially derivative. His early guild socialism had been remarkably innocent of any serious economic theory at all.

Yet, instead of confirming Cole as of only minor importance, what this really serves to emphasize is the remarkable nature of his contribution to practical economics between the wars. If his economic theory was derivative, he derived it from sources that enabled him to construct radical policy proposals to combat slump and unemployment. Drawing particularly upon the 'underconsumption' (or 'over-saving') analysis of capitalism developed by J.A. Hobson, Cole mounted a sustained critique of economic orthodoxy in relation to unemployment throughout the 1920s and argued the need for demand stimulation and a bold programme of public works and investment. The great merit of Hobsonian economic theory for a socialist like Cole was that it provided the materials from which capitalism could be both indicted and reformed.

Cole's recovery programme remained substantially the same in the 1930s, but from the early years of that decade he displayed a clearer understanding of how such a programme was to be financed. His policy proposals were already proto-Keynesian, but from the early 1930s (when he worked with Keynes on the Economic Advisory Council) he analysed the economic situation from a recognizably Keynesian perspective. Reviewing Keynes's *General Theory* in the *New Statesman* (the house magazine of the British Left), Cole described it as 'the most important theoretical economic writing since Marx's *Capital*, or, if only classical economics is to considered

as comparable, since Ricardo's *Principles*'. Above all, it provided the theoretical credentials for his own dissenting economics.

However, if Keynes had to be absorbed by the Left, and mobilized for a recovery programme, Cole also took the view that it was necessary to look beyond the conditions of short-term stabilization and towards the development of a 'new' economics of socialism. He therefore emerged as a leading advocate of socialist economic planning in the 1930s, but for the rest of his life (and after the war from the vantage point of the Chichele Chair of Social Political Theory at Oxford) he continued to search for a form of socialist economy consistent with his prior commitment to a form of non-bureaucratic socialist democracy.

Selected Works

1929. *The next ten years in British social and economic policy*. London: Macmillan.
1932. *Economic tracts for the times*. London: Macmillan.
1935. *Principles of economic planning*. London: Macmillan.
1950. *Socialist economics*. London: Gollancz.

References

Cole, M. 1971. *The life of G.D.H. Cole*. London: Macmillan.
Wright, A.W. 1979. *G.D.H. Cole and socialist democracy*. Oxford: Clarendon.

Collective Action

Mancur Olson

Abstract

The logic of collective action undermines the assumption that common interests are always promoted by their beneficiaries. Where the number of beneficiaries is large, the benefits of collective action are a public good: beneficiaries will gain whether or not they participate in promoting them, while their individual efforts cannot secure them. Small groups can use selective incentives to ensure that their members contribute to promoting their common interests. This typically results in the paradoxical 'exploitation of the great by the small'. The logic of collective action helps explain many notable examples of economic growth and stagnation since the Middle Ages.

Keywords

Anarchy; Bargaining; Cartels; Class conflict; Collective action; Collective bargaining; Collusion; Common interests; Countervailing power; Encompassing organizations; Excess burden; Exploitation; Galbraith, K.; Group theory; Industrial revolution; Invisible hand; Latent groups; Lobbying; Mercantilism; Non-excludability; Olson, M.; Patronage dividends; Public choice; Public goods; Revelation of preferences; Samuelson, P.; Selective incentives; Smith, A.; Strategic behaviour; Technical progress; Trade unions; Wicksell, J.

JEL Classifications

D71

For a long while, economists, like specialists in other fields, often took it for granted that groups of individuals with common interests tended to act to further those common interests, much as individuals might be expected to further their own interests. If a group of rational and self-interested individuals realized that they would gain from political action of a particular kind, they could be expected to engage in such action; if a group of workers would gain from collective bargaining, they could be expected to organize a trade union; if a group of firms in an industry would profit by colluding to achieve a monopoly price, they would tend to do so; if the middle class or any other class in a country had the power to dominate, that class would strive to control the government and run the country in its own interest. The idea that there was some tendency for groups to

act in their common interests was often merely taken for granted, but in some cases it played a central conceptual role, as in some early American theories of labour unions, in the 'group theory' of the 'pluralists' in political science, in J.K. Galbraith's concept of 'countervailing power', and in the Marxian theory of class conflict.

More recently, the explicit analysis of the logic of individual optimization in groups with common interests has led to a dramatically different view of collective action. If the individuals in some group really do share a common interest, the furtherance of that common interest will automatically benefit each individual in the group, whether or not he has borne any of the costs of collective action to further the common interest. Thus the existence of a common interest need not provide any incentive for individual action in the group interest. If the farmers who grow a given crop have a common interest in a tariff that limits the imports and raises the price of that commodity, it does not follow that it is rational for an individual farmer to pay dues to a farm organization working for such a tariff, for the farmer would get the benefit of such a tariff whether he had paid dues to the farm organization or not, and his dues alone would be most unlikely to determine whether or not the tariff passed. The higher price or wage that results from collective action to restrict the supply in a market is similarly available to any firm or worker that remains in that market, whether or not that firm or worker participated in the output restriction or other sacrifices that obtained the higher price or wage. Similarly, any gains to the capitalist class or to the working class from a government that runs a country in the interests of that class, will accrue to an individual in the class in question whether or not that individual has borne the costs of any collective action. This, in combination with the extreme improbability that a given individual's actions will determine whether his group or class wins or loses, entails that a typical individual, if rational and self-interested, would not engage in collective action in the interest of any large group or class.

Analytically speaking, the benefits of collective action in the interest of a group with a common interest are a public or collective good to that group; they are like the public goods of law and order, defence, and pollution abatement in that voluntary and spontaneous market mechanisms will not provide them. The fundamental reality that unifies the theory of public goods with the more general logic of collective action is that ordinary market or voluntary action fails to obtain the objective in question. It fails because the benefits of collective or public goods, whether provided by governments or non-governmental associations, are not subject to exclusion; if they are received by one individual in some group, they automatically also go to the others in that group (Olson 1965).

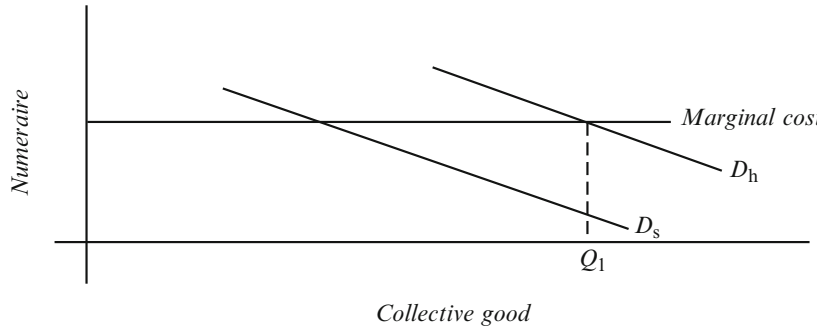
Since many groups with common interests obviously do not have the power to tax or any comparable resource, the foregoing logic leads to the prediction that many groups that would gain from collective action will not in fact be organized to act in their common interests. This prediction is widely supported. Consumers have a common interest in opposing the legislation that gives various producer groups supra-competitive prices, and they would sometimes also have a common interest in buyers' coalitions that would countervail producer monopolies, but there is no major country where most consumers are members of any organization that works predominantly in the interest of consumers. The unemployed similarly share a common interest, but they are nowhere organized for collective action. Neither do most taxpayers, nor most of the poor, belong to organizations that act in their common interest (Austen-Smith 1981; Brock and Magee 1978; Chubb 1983; Hardin 1982; Moe 1980; Olson 1965).

Though some groups can never act collectively in their common interest, certain other groups can, if they have ingenious leadership, overcome the difficulties of collective action, though this usually takes quite some time. There are two conditions either of which is ultimately sufficient to make collective action possible. One condition is that the number of individuals or firms that would need to act collectively to further the common interest is sufficiently small; the other is that the groups should have access to 'selective incentives'.

The way that small numbers can make collective action possible at times is most easily evident on the assumption that the individuals in a group with a common interest are identical. Suppose there are only two large firms in an industry and that each of these firms will gain equally from any government subsidy or tax loophole for the industry, or from any supra-competitive price for its output. Clearly each firm will tend to get the benefit of any lobbying it does on behalf of the industry, and this can provide an incentive for some unilateral action on behalf of the industry. Since each firm's action will have an obvious impact on the profits of the other, the firms will have an incentive to interact strategically with and bargain with one another. There would be an incentive to continue this strategic interaction or bargaining until a joint maximization or 'group optimal' outcome had been achieved. This same logic obviously also applies to collective action in the form of collusion to obtain a supra-competitive price, and thus we obtain the well-known incentive for oligopolistic collusion in concentrated industries whenever there are significant obstacles to or costs of entry. As the number in a group increases, however, the incentive to act collectively diminishes; if there are ten identical members of a group with a common interest, each gets a tenth of the benefit of unilateral action in the common interest of the group, and if there are a million, each gets one millionth. In this last case, even if there were some incentive to act in the common interest, that incentive would cease long before a group-optimal amount of collective action had taken place. Strategic interaction or voluntary bargaining will not occur since no two individuals have an incentive to interact strategically or to bargain with one another. This is because the failure of one individual to support collective action will not them have any perceptible effect on the incentive any other individual faces so there is no incentive for strategic interaction or rational bargaining. Thus we obtain the result that, in time, sufficiently small groups can act collectively, but that this incentive for collective action decreases monotonically as the group gets larger and disappears entirely in sufficiently large or 'latent' groups.

When the parties that would profit from collective action have very different demand curves, the party with the highest absolute demand for collective action will have an incentive to engage in some amount of collective action when no other member of the group has such an interest. This leads to a paradoxical 'exploitation of the great by the small'. This is true to a greater degree and is evident much more simply if income effects are ignored, as in the demand curves for a collective good depicted in the figure below. When the party with the highest demand curve for the collective good, D_h, has obtained the amount of the collective good, Q_1, that is in its interest unilaterally to provide, any and all parties with a lower demand curve, such as D_s, will automatically receive this same amount, and thus have no incentive to provide any amount at all! (Olson 1965). When income effects and certain 'private good' aspects of some collective goods are taken into account the results are less extreme, but a distribution of burdens disproportionality unfavourable to the parties with the absolutely larger demands tends to remain. This disproportion has been evident, for example, in various military alliances and international organizations, in cartels, and in metropolitan areas in which metropolis-wide collective goods are provided by independent municipalities of greatly different size (Olson and Zeckhauser 1966; Sandler 1980) Fig. 1.

The other condition, besides small numbers, that can make collective action possible, is 'selective incentives'. Those large groups that have been organized for collective action for any substantial period of time are regularly found to have worked out special devices, or selective incentives, that are functionally equivalent to the taxes that enable governments to provide public goods (Olson 1965; Hardin 1982). These selective incentives either punish or reward individuals depending on whether or not they have borne a share of the costs of collective action, and thus give the individual an incentive to contribute to collective action that no good that is or would be available to all could provide. The most obvious devices of this kind are the 'closed shop' and picket line arrangements of labour unions, which often make union membership a condition of



Collective Action, Fig. 1

employment and control the supply of labour during strikes (see, for example, McDonald 1969; Gamson 1975). Upon investigation it becomes clear that labour unions are not in this respect fundamentally different from other large organizations for collective action, which regularly have selective incentives that, though usually less conspicuous than the closed shop or the picket line, serve the same function.

Farm organizations in several countries, and quite notably in the United States, obtain most of their membership by deducting the dues in farm organizations from the 'patronage dividends' or rebates of farm cooperatives and insurance companies that are associated with the farm organizations. The professional associations representing such groups as physicians and lawyers characteristically have either relatively discreet forms of compulsion (such as the 'closed bar') or subtle individual rewards to association members, such as access to professional publications, certification, referrals, and insurance. In small groups, and sometimes in large 'federal' groups that are composed of many small groups, social pressure and social rewards are also important sources of selective incentives.

The selective incentives that are needed if large groups are to organize for collective action are less often available to potential entrants or those at the lower levels of the social order than to established and well-placed groups. The unemployed, for example, obviously do not have the option of making membership of an organization working in their interest a condition of employment, nor do they naturally congregate as the employed do at workplaces where picket lines may be established. Those who would profit from entering a cartelized industry or profession are similarly almost always without selective incentives. Experience in a variety of countries also confirms that those with higher levels of education and skill have better access to selective incentives than lower income workers; highly trained professionals such as physicians and attorneys usually come to be well organized before labour unions emerge, and the unions of skilled workers normally emerge before unions representing less skilled workers. The correlation between income and established status and access to selective incentives works in the same direction as the lesser difficulty of collective action of small groups of large firms in relatively concentrated industries explained above. Together these two factors generate a tendency for collective action to have, in the aggregate though not in all cases, a strong anti-egalitarian and pro-establishment impact (Olson 1984).

The study of collective action goes back to the beginnings of economics, but then came to be strangely neglected during most of the rest of the history of the subject. Though this is not generally realized, the study of collective action, admittedly only in an inductive and intuitive way, was a crucial part of Adam Smith's analysis of the inefficiencies and inequities in the economies he observed (Smith 1776). Smith even noted that the main beneficiaries of collective action in his time were by no means the poor or those of average means. He also emphasized the tendency for urban interests to profit from collective action at the expense of rural people, because the

geographical dispersion of agricultural interests areas made it more difficult for them to combine to exert political influence or to fix prices; this emphasis presumably owed something to the poor transportation and communication systems in his day, which presumably obstructed the organization of rural interests more in his time than it does in developed countries now.

The label that Adam Smith gave to the set of public policies, monopolistic combinations, and ideas that he attacked was, after all, 'mercantilism', because the single most important source of the evils was the collective action of merchants, or merchants and 'masters', especially those organized into guilds or 'corporations'. In his discussions of the 'Inequalities Occasioned by the Policy of Europe' and of 'The Rent of Land' (Bk. I, ch. 10, pt. ii and ch. 11), Smith emphasized that 'whenever the legislature attempts to regulate the differences between masters and their workmen, its counsellors are always the Masters'. Similarly,

> it is everywhere much easier for a rich merchant to obtain the privilege of trading in a town corporate, than for a poor artificer to obtain that of working in it . . . Though the interest of the labourer is strictly connected with that of the society . . . his voice is little heard and less regarded.

The rural interests are similarly at a disadvantage, according to Smith, especially as compared with those in 'trade and manufacturers':

> The inhabitants of a town, being collected into one place, can easily combine together. The most insignificant trades carried on in towns have accordingly, in some place or another, been incorporated . . . voluntary associations and agreements prevent that free competition which they cannot prohibit. . . .The trades which employ but a small number of hands run most easily into such combinations. . . . People of the same trade seldom meet together, even for merriment and diversion, but the conversation ends in a conspiracy against the public, or in some contrivance to raise prices.

By contrast, 'the inhabitants of the country, dispersed in distant places, cannot easily combine together'.

These passages, though not in the order they appear in Smith, nonetheless correctly convey his alertness to collective action. Though the handicap that rural interests face in organizing for collective action is far less in developed countries today than it was in Smith's time, even this part of his argument still generally holds true in the developing countries, where transportation and communication in the rural areas are poor, peasants are generally unrepresented, and agricultural commodities normally underpriced (Anderson and Hayami 1986; Schultz 1978; Olson 1985).

Adam Smith's insights into collective action and its consequences were ignored until recent times. Presumably one reason is that most economists in the 19th and early 20th centuries were mainly interested in the logic of the case for competitive markets. The logic of collective action, by contrast, is really a general statement of the logic of market failure; it embodies the central insight of the theories of public goods and externalities, that markets and voluntary market-type arrangements do not generally work in those cases where the beneficiaries of any collective good or benefit cannot be excluded because they have not paid any purchase price or dues (Baumol 1952). It was not until Knut Wicksell's New Principle of Just Taxation' was published in German in 1896 (Musgrave and Peacock 1967) that any economist revealed a clear understanding of the nature of public goods, and only with the publication of Samuelson's articles in the 1950s (Samuelson 1955) that this idea came to be generally understood in the English-speaking world.

A second obstacle to the development of the logic of collective action was that collective action by governments was normally taken for granted. Notwithstanding the difficulties of collective action, anarchy is relatively rare because a government that provides some sort of law and order quickly takes over. This in turn is due to conquerors and the gains they obtain in increased tax revenues from establishing some system of law and order and property rights. In the absence of the provision of these most elemental collective goods, there is not much for a conqueror to take, so the historic first movement of the invisible hand is evident in the incentive conquerors have to establish law and order. Those who lead the governments that succeed conquerors obviously must

maintain a system of law and order if they are to continue collecting significant tax revenues. Since governments providing basic collective goods have been ubiquitous, the classic writers on public goods like Wicksell and Samuelson did not even ask how collective goods emerged in the first place. They focused instead on how to determine what was an appropriate sharing of the tax burdens and on the difficulty of determining what level of provision of public goods was Pareto-optimal. This in turn naturally led to Wicksell's recommendation that only those public expenditures that could, with an approximate allocation of the tax burdens, command approximate unanimity, should normally be permitted, and to Samuelson's and Musgrave's (1959) concern for the non-revelation of preferences for public goods. The difficulties of collective action and public good provision on a voluntary basis therefore naturally did not gain any theoretical attention.

When, as in the new political economy or public choice, the focus is also on the efforts of extra-governmental groups to obtain the gains from lobbying, cartelization, and collusion, and on private action to obtain collective benefits of other kinds, a more general conception becomes natural (Barry and Hardin 1982; Olson 1965; Taylor 1976). It then becomes clear that the likelihood of voluntary collective action depends dramatically on the size of the group that would gain from collective action. When a group is sufficiently small and there is time for the needed bargaining, the desired collective goods will normally be obtained through voluntary cooperation (Frohlich et al. 1971). If there are substantial differences in the demands for the collective good at issue, there will be the aforementioned paradoxical 'exploitation of the great by the small'. When the number of beneficiaries of collective action is very large, voluntary and straightforward collective action is out of the question, and taxes or other selective incentives are indispensable. Selective incentives are available only to a subset of those extra-governmental groups that would gain from collective action. Even those extra-governmental groups that do have the potential of organizing through selective incentives will usually have great difficulty in working out these (often subtle) devices, and will normally succeed in overcoming the great difficulties of collective action only when they have relatively ingenious leadership and favourable circumstances.

If follows that it is only in long-stable societies that many extra-governmental organizations for collective action will exist. In societies where totalitarian repression, revolutionary upheavals, or unconditional defeat have lately destroyed organizations for collective action, few groups will have been able in the time available to have overcome the formidable difficulties of collective action. It has been shown elsewhere (Mueller 1983; Olson 1982), that (unless they are very 'encompassing') organizations for collective action have extraordinarily anti-social incentives; they engage in distributional struggles, even when the excess burden of such struggles is very great, rather than in production. They also will tend to make decisions slowly and thereby retard technological advance and adaptations to macroeconomic and monetary shocks. It follows that societies that have been through catastrophes that have destroyed organizations for collective action, such as Germany, Japan, and Italy, can be expected to enjoy 'economic miracles'. An understanding of collective action also makes it possible to understand how Great Britain, the country that with industrial revolution discovered modern economic growth and had for nearly a century the world's fastest rate of economic growth, could by now have fallen victim to the 'British disease'. The logic of collective action, in combination with other theories, also makes it possible to understand many of the other most notable examples of economic growth and stagnation since the Middle Ages, and also certain features of macroeconomic experience that contradict Keynesian, monetarist, and new classical macroeconomic theories (Balassa and Giersch 1986).

See Also

► Bargaining
► Collective Action (New Perspectives)

- ▶ Public Choice
- ▶ Social Choice

Bibliography

Anderson, K., and Y. Hayami. 1986. *The political economy of protection*. Sydney: George Allen & Unwin.

Austen-Smith, D. 1981. Voluntary pressure groups. *Economica* 48: 143–153.

Balassa, B., and H. Giersch (eds.). 1986. *Economic incentives.* Proceedings of the international economic association, London: Macmillan.

Barry, B., and R. Hardin (eds.). 1982. *Rational man and irrational society.* Beverly Hills: Sage.

Baumol, W.J. 1952. *Welfare economics and the theory of the state*. Cambridge, MA: Harvard University Press.

Brock, W., and S. Magee. 1978. The economics of special interest groups: The case of the tariff. *American Economic Review* 68: 246–250.

Chubb, J. 1983. *Interest groups and bureaucracy.* Stanford: Stanford University Press.

Frohlich, N., J. Oppenheimer, and O. Young. 1971. *Political leadership and collective boards*. Princeton: Princeton University Press.

Gamson, W.A. 1975. *The strategy of social protest*. Homewood: Dorsey.

Hardin, R. 1982. *Collective action*. Baltimore: Johns Hopkins University Press for Resources for the Future.

McDonald, D.J. 1969. *Union man*. New York: Dutton.

Moe, T.M. 1980. *The organization of interests*. Chicago: University of Chicago Press.

Mueller, D.C., (ed.). 1983. *The political economy of growth.* New Haven: Yale University Press.

Musgrave, R.A. 1959. *The theory of public finance*. New York: McGraw-Hill.

Musgrave, R.A., and A.T. Peacock (eds.). 1967. *Classics in the theory of public finance,* 2nd ed. New York: McGraw-Hill.

Olson, M.L. 1965. *The logic of collective action*. Cambridge, MA: Harvard University Press.

Olson, M.L. 1982. *The rise and decline of nations*. New Haven: Yale University Press.

Olson, M.L. 1984. Ideology and growth. In *The legacy of reaganomics*, ed. C.R. Hulten and I.V. Sawhill. Washington, DC: Urban Institute Press.

Olson, M.L. 1985. Space, organization, and agriculture. *American Journal of Agricultural Economics* 67: 928–937.

Olson, M.L., and R. Zeckhauser. 1966. An economic theory of alliances. *The Review of Economics and Statistics* 48: 266–279.

Samuelson, P.A. 1955. Diagrammatic exposition of a theory of public expenditure. *The Review of Economics and Statistics* 37: 350–356.

Sandler, T. (ed.). 1980. *The theory and structure of international political economy.* Boulder: Westview.

Schultz, T.W. 1978. *Distortion of agricultural incentives*. Bloomington: Indiana University Press.

Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations*. London: J.M. Dent. 1910.

Taylor, M. 1976. *Anarchy and cooperation*. London: John Wiley.

Wicksell, K. 1896. A new principle of just taxation. Trans. from the German by J.A. Buchanan, in Musgrave and Peacock (1967).

C

Collective Action (New Perspectives)

David P. Myatt

Abstract

Olson's logic of collective action predicts that public-good provision is most likely to fail when the size of the consumer group is large; his public goods are partially rival, and so the private cost of provision is relatively high. With a pure public good, this logic no longer applies, and so attention turns to producer groups. When provision involves teamwork (so that the collective action succeeds when everyone works together) then coordination problems arise. Modern techniques suggest that 'good' equilibria in which provision is successful are robust only when the costs of provision fall below private rather than social benefits.

Keywords

Chicken games; Collective action; Cournot contributions games; Critical mass theory; Equilibrium-selection problem; Externalities; Global games; Interdependent consumption; Interdependent production; Market failure; Multiple equilibria; Olson, M.; Provision games; Public goods; Risk dominance; Schelling, T.; Selective incentives; Strategic voting; Strategy revision; Teamwork dilemma; Volunteer's dilemma

JEL classifications

D71

In a review conducted on behalf of the UK Government, Stern (2007) concluded that 'climate change is a serious global threat, and demands an urgent global response ... the benefits of strong and early action far outweigh the economic costs of not acting'. The cuts in emissions that he suggested could generate global benefits. However, the costs would be borne individually by those making significant cuts (developed nations) or by those sacrificing future opportunities (rapidly developing nations).

A shared desire to cut greenhouse-gas emissions generates a classic problem of collective action: a group with common interests must rely on voluntary individual optimization for the pursuit of those interests. Stern's 'urgent global response' to a 'serious global threat' requires nations to act. Such sovereign states need respond only to their own incentives; any participation is voluntary. Within each state, the pursuit of national objectives is not automatic; environmental effects stem from the decisions of individual agents. Even if it were in a state's collective interest to support a collective action against climate change, it cannot be assumed that constituents of that state would individually offer their backing.

To economists, the collective-action problem boils down to the private provision of a public good or the private exploitation of a common resource. Law and order, military defence and pollution control are classic textbook examples of public goods: the benefits of provision are non-excludable, and so private providers fail to capture the full impact of their contributions. This market failure leads to inefficient under-provision. On the other hand, the commons exploitation of traffic congestion and commercial fishing yield negative externalities: market failure yields to inefficient overindulgence in these activities. In both cases, individuals fail to pursue efficiently their collective objectives.

The idea that group members will not always pursue their common interests was once not accepted widely. In his article in the first edition of the *New Palgrave*, Mancur Olson (1987) observed that 'economists, like specialists in other fields, often took it for granted that groups of individuals with common interests tended to act to further those common interests, much as individuals might be expected to further their own interests'. He persuasively argued that 'the existence of a common interest need not provide any incentive for an individual action in the group interest'. Hence consumers may fail to campaign for their collective protection, unions may fail to protect all their members, oligopolists may fail to maintain collusive prices, and nations of the world may fail to prevent further climate change.

Olson's point was simple and is now familiar: when contemplating choice, individuals consider only the private impact of their actions. For the classic case of a public good, an individual faces the full marginal cost of provision but fails to account for the benefit spilling over to others; the presence of positive externalities leads to under-provision. If an individual could internalize these externalities, perhaps by excluding the consumption of others and charging them for it, then efficiency could be restored. Alas, pure public goods are non-excludable, and hence this route to efficiency is blocked.

Nevertheless, as long as individuals enjoy some private benefit from voluntary action then we can expect some, albeit too little, provision of public goods. The extent of any inefficiency depends upon the nature of the collective-action problem, the availability of mechanisms to restore efficiency, and the size and nature of the relevant group. Olson (1965) concluded that 'unless the number of individuals in a group is quite small, or unless there is coercion or some other special device to make individuals act in their common interest, rational, self-interested individuals will not act to achieve their common or group interests'. In the context of small groups, when partial provision is deemed possible, he identified 'a surprising tendency for the 'exploitation' of the great by the small'. These claims led to his theory of groups: (*a*) collective actions fail when the groups are large; (*b*) larger factions bear a disproportionate share of any provision; and (*c*) selective incentives are necessary if groups are to succeed. These three claims are considered in turn, before attention turns to a rather different perspective on collective action.

The first claim is Olson's 'group size' hypothesis: private provision should fall as a group

grows larger. Olson (1965) painted a picture of a meeting at which too few people make careful contributions: 'When the number of participants is large, the typical participant will know that his own efforts will probably not make much difference to the outcome, and that he will be affected by the meeting's decision in much the same way no matter how much or how little effort he puts into studying the issues.' More directly, the claim is that the private benefit of any voluntary contribution falls with the group's size; equivalently, the private cost for any particular level of public provision rises with the group size. This claim leans on two implicit assumptions. First, an increase in the number consuming the good leads to an increase in the provision cost, and hence the good is (at least partially) rival; it is an impure public good. Second, the group size corresponds to the number of consumers, and not to the size of the contributor pool.

These two implicit assumptions that underpin the group-size hypothesis are often valid. For instance, the global climate change that worried Stern (2007) corresponds to a 'large group' global collective-action problem (Sandler 2004). Nevertheless, the assumptions often exclude interesting collective-action problems. The first assumption rules out pure public goods. Consider, for instance, the contemporary voluntary provision of open-source software (Raymond 1998; Johnson 2002; Lerner and Tirole 2002). The typical license under which such software is distributed 'requires that the source code ... be made available to everyone, and that the modifications made by its users also be turned back to the community' (Lerner and Tirole 2001). This a modern instance of the 'collective invention' documented by Allen (1983). Open-source software is automatically non-excludable. Of course, software is a classic instance of a non-rival good: consumption by one individual does not hamper the consumption opportunities of others. Hence, an increase in the size of the group consuming the good, while fixing the size of the group able to provide it, has no direct impact on incentives.

Olson's second claim was that provision costs fall on larger members of a group. The idea is that such members consume large shares of the public good, and so face a relatively large private benefit. Once again, this builds upon the assumption that the collective output is rival; for a pure public good, the same logic would predict that those who care most contribute most, and such contributors need not be large in a conventional sense.

Olson's third claim concerned the possible response to the problem of collective action. Such a response requires, according to this claim, selective incentives that are 'functionally equivalent to the taxes that enable governments to provide public goods ... [they] either punish or reward individuals depending on whether or not they have borne a share of the costs of collective action, and thus give the individual an incentive to contribute ...' (Olson 1987). The classic example of selective incentives is the 'closed shop' of labour unions; to enjoy the benefits of collective union bargaining power each worker must be a member, and hence pay the costs of any strike action. Interestingly, when the selective incentive is based on preventing a group member from enjoying the collective output then the implicit assumption is that the public good is at least partially excludable.

In summary, Olson (1965, 1987) forcefully clarified the inescapable logic of collective action: any theory of group behaviour must rely upon the incentives faced by individuals, and not simply assume that groups pursue their common interests. His theory of groups remains relevant for many contemporary problems. However, it steps outside the world of pure public goods by assuming the interdependent consumption of an impure public good, and does not allow for interdependence of production. Put more succinctly, Olson's groups consist of public good consumers rather than public good providers.

Attention now refocuses on collective-action problems in which economic players non-cooperatively choose whether to participate in the private production of a pure public good. Crucially, there can be interdependence of production: the incentive to participate in a collective action depends on the expected participation of others. Decisions become genuinely strategic, and this changes the nature of the collective-action problem.

A little notation proves helpful. Amongst n players, write x_i for the action of player i, and collect the actions of everyone together into a vector x. Payoffs satisfying

$$u_i(x) = G(x) - c_i(x_i) \quad (1)$$

comprise the value $G(x)$ of public good and the private cost $c_i(x_i)$ that player i incurs when contributing to it; the externality imposed on others is captured by $(n-1)G(x)$. The nature of the strategic interaction amongst players depends upon the form taken by $G(x)$. A simple specification is when x_i is a positive real number and $G(x) \equiv \sum_{i=l}^{n} x_i$. A player's decision is strategically independent of others' actions: he simply equates the private marginal benefit of the public good to its private marginal cost via the first-order condition $1 = c'(x_i)$, yielding the usual under-provision problem (Cornes and Sandler 1996).

A second natural specification to consider is where $G(x) \equiv F\left(\sum_{i=l}^{n} x_i\right)$ for some nicely behaved concave production function $F(\cdot)$. This falls within the class of Cournot contributions games (Chamberlin 1974; McGuire 1974; Young 1982; Cornes and Sandler 1985; Bergstrom et al. 1986; Bernheim 1986). Here, strategic interaction is non-trivial since the marginal benefit of increased public good provision depends on the total contributions of all players. Nevertheless, a unique Nash equilibrium involves under-provision. The associated literature concerned itself with the comparative-static properties of such models, including the response of public good output and the burden of provision to the redistribution of income (Warr 1983; Kemp 1984).

These first two examples of eq. (1) simply flesh out the implicit model of Olson (1965). The nature of the collective action problem changes significantly when $G(x)$ takes on more interesting and yet plausible shapes. For instance, $G(\cdot)$ might take a weakest link ($G(x) = \min\{x_i\}$) or best shot ($G(x) = \min\{x_i\}$) form (Hirshleifer 1983, 1985); these are special cases of symmetric but non-additive specifications (Cornes 1993).

Here, however, attention turns to situations in which the success of a collective action (that is, the successful provision of a public good) turns upon either the participation of a critical mass of players, or contributions that exceed a particular threshold. Returning once more to the economics of climate change, a plausible scenario is one in which the ice caps melt unless carbon emissions are pushed down below a critical level. Whereas in a Cournot contributions game the incentive to contribute decreases with the participation of others, here it may increase: a nation may find it worthwhile to chase environmental targets if and only if it expects others to play their part in international agreements.

A central feature of threshold-based scenarios is that an individual's decision depends on aggregate participation. This is easiest to explore in a binary-action game where $x_i \in \{0,1\}$ for each player i; hence $x_i = 1$ can be interpreted as individual participation in a collective action. In many such situations, the incentive to participate depends on the number of others who do so. Hence, writing $\Delta u_i(x)$ for this incentive,

$$\Delta u_i(x) \equiv P(m) \text{ where } m = \sum_{j \neq i} x_j. \quad (2)$$

When $P(m) < 0$ for all m, no players participate; this is an n-player Prisoner's Dilemma. If $P(m)$ decreases with m, then the unique equilibrium entails the participation of m^* players, where $P(m^* - 1) > 0 > P(m^*)$; for the Cournot games considered above the participation m^* might be socially suboptimal. If $P(m)$ increases with m, so that there is a threshold m^* satisfying $P(m^* - 1) < 0 < P(m^*)$, then there are two pure-strategy Nash equilibria, one in which everyone participates, and one in which the collective action fails. This means that the problem of collective action becomes one of coordination.

Games satisfying eq. (2) drew the insightful attention of Schelling (1973, 1978). He opened his analysis by describing the use of protective helmets in ice hockey: players were willing to wear helmets only if others did so too. Other sociological examples are easy to find: members of a crowd will join a protest only if others do so (Berk 1974; Granovetter 1978) and successful consumer boycotts require a critical mass (Innes 2006).

Political situations can also fit eq. (2). Consider a plurality rule election in which a group wishes to prevent the success of a disliked incumbent candidate. They can do so if and only if a critical number m^* abandon their first-preference candidate and vote for their second choice. Setting $P(m^*-1) > 0$ and $P(m) < 0$ otherwise yields a strategic-voting model (Palfrey 1989; Myerson and Weber 1993; Cox 1994, 1997; Myatt 2007).

In sociology, collective-action games with threshold properties fall under the umbrella of the theory of critical mass (Oliver et al. 1985; Oliver and Marwell 1988; Marwell et al. 1988; Marwell and Oliver 1993). Alas, these sociologists had no theoretical machinery for selecting between multiple equilibria. In economics, multiple equilibria arise in threshold-driven step-level public goods games (Palfrey and Rosenthal 1984). Once again, the problem of coordination boils down to a need to choose amongst multiple equilibria. Fortunately, recent contributions to economics allow some progress to be made on the equilibrium-selection problem.

To explore further, it is instructive to consider a simple world: two individuals (A and B) either participate (Y) or not (N) in a collective action. Participationinvolves a private cost (either c_A or c_B), but may provide a public good to be enjoyed by both players. A natural representation is via a simple 2×2 strategic form game (Fig. 1).

In the 'provision game' a participant produces a public good worth v to everyone. A player's marginal product is strategically independent: the incentive for player A to participate is always $v - c_A$, and hence he does so if and only if $v > c_A$. However, this generates a spillover of v for player B, and hence the social gain is $2v - c_A$. The parameter configuration $2v > c_A > v$ yields the classic under-provision of a public good.

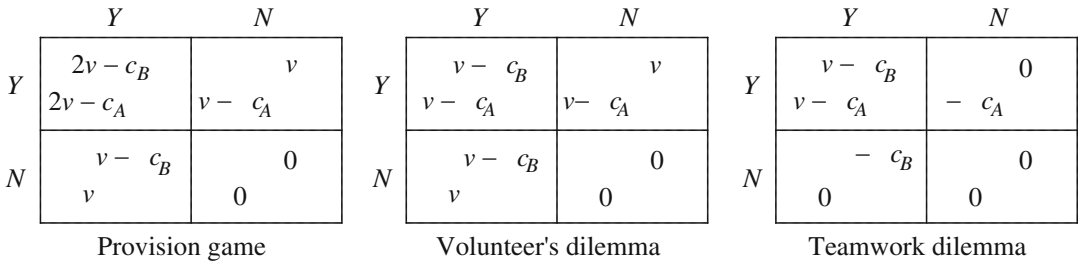
But what if there is strategic interdependence? Suppose that only one player need provide, so that a second participant generates a cost but no additional benefit. This 'volunteer's dilemma' (Diekmann 1985) is a textbook game of 'chicken' (Lipnowski and Maital 1983). If $2v > c_A > v$ and $2v > c_B > v$ then neither player is willing to participate even though it is socially optimal for someone to do so. However, if $v > c_A$ then player A participates so long as player B does not. If $v > c_B$, then there are two pure-strategy Nash equilibria in which a single player provides the public good. But who provides?

One possibility is to use risk dominance (Harsanyi and Selten 1988) as a selection criterion. The risk-dominant equilibrium is that which maximizes the product of players' incentives to remain at the equilibrium. So, in the volunteer's dilemma, the equilibrium in which A provides is risk-dominant if $(v - c_A)c_B > (v - c_B)c_A$, which holds if and only if $c_A < c_B$: the most efficient provider volunteers. Following Olson (1965), the strong (low-cost providers) bear the cost of provision to the benefit of the weak.

A coordination problem also arises in the 'teamwork dilemma' (Fig. 1) where both players are needed for the collective action to succeed. This is an assurance or 'stag hunt' game: as long as $v > c_A$ and $v > c_B$ there is a pure-strategy equilibrium in which both players participate, and a second with no participation in which the collective action fails. The former equilibrium is risk dominant if and only if $(v - c_A)(v - c_B) > c_A c_B$, which boils down to $v > c_A + c_B$; this requires a single private (not social) benefit from the public good to exceed the total private cost of provision. If $2v > c_A + c_B > v$, then the collective action fails even though it would be socially optimal for it to succeed. Once again, this is a return to Olson (1965): success of the collective action relies on private incentives.

All well and good, but can the criterion of risk dominance be justified? In the recent literature two contrasting approaches lead to the same answer.

The theory of global games (Carlsson and Van Damme 1993; Morris and Shin 2003) supposes that players do not share common knowledge of the payoffs of games. Instead, players must rely upon privately observed signals of the game being played. For instance, players may be unsure of the true value v of the public good, and see an estimate of it. Crucially, this estimate allows them to infer not only this value but also the probable signals received by others, and hence their opponents' likely behaviour. When signals become very precise then the play of a simple 2×2 game almost



Collective Action (New Perspectives), Fig. 1 Public-good provision games

always coincides with the risk-dominant Nash equilibrium (Carlsson and Van Damme 1993).

Others have selected equilibria by studying the evolving play. Players (or populations from which players are drawn) may adjust their play over time in the direction of myopic best-reply, but occasionally 'mutate' to a different strategy (Kandori et al. 1993; Young 1993, 1998). As the probability of mutations vanishes, play in the long run focuses on a single stochastically stable equilibrium (Foster and Young 1990). In a symmetric teamwork dilemma, it picks out the risk dominant equilibrium.

Can modern literature say anything about the general case of eq. (1)? Players act as though they are attempting to maximize jointly the single real-valued function

$$\rho(x) \equiv G(x) - \sum_{i=l}^{n} c_i(x_i). \qquad (3)$$

This is a potential function, and yields a potential game (Monderer and Shapley 1996). This function has a natural interpretation: the private benefit that a single individual derives from a public good, minus the total private costs involved in its provision.

Clean results emerge when play of a potential game evolves via a payoff-responsive stochastic strategy-revision process (Blume 1993, 1995, 1997; Brock and Durlauf 2001; Blume and Durlauf 2001, 2003). Over time, players occasionally revise their strategies. When a player does so, his decision is not a myopic best reply to the current play of others, but rather a quantal response (McKelvey and Palfrey 1995): the log odds of choosing one action rather than another is determined by the difference in their payoffs, and so he is more likely to choose better performing strategies. An inspection of eq. (3) reveals that the difference in a player's payoffs is equal to the difference in potential; the potential function captures the essential strategic interaction of the game.

Allowing play to evolve, the strategy-revision process is drawn towards the states-of-play with the highest potential. In the long run, when quantal responses approximate best replies, the process spends almost all time in the state that maximizes $\rho(x)$: evolution leads players to maximize the difference between a single private benefit and total private costs rather than social welfare which would incorporate the full social benefit of $nG(x)$.

This approach can be applied to the teamwork dilemma: the potential of the state-of-play in which neither player participates is zero, and the potential of the equilibrium in which the collective action succeeds is $v - (c_A + c_B)$. The latter equilibrium has positive potential if and only if $v > c_A + c_B$: only if a private individual would be willing to step forward and pay the full cost of provision himself will the collective action succeed. So, whereas it may at first appear that the success of a collective action (the coordinated play of f*Y*,*Y*g in the teamwork dilemma) can follow from the interdependence of team members, evolving play results in failure (the play of f*N*,*N*g in the teamwork dilemma) unless a private individual would be willing to fund the collective action himself.

On reflection, this should be unsurprising. Each step of evolving play (or each step of

reasoning in a global-game argument) is driven by reference to private incentives. So what lesson should be taken away? Even when a group's problem is one of coordination, its members cannot escape Olson's (1965, 1987) fundamental logic of collective action.

See Also

- ▶ Collective Action
- ▶ Externalities
- ▶ Public Goods

Bibliography

Allen, R.C. 1983. Collective invention. *Journal of Economic Behavior and Organization* 4: 1–24.

Bergstrom, T.C., L. Blume, and H.R. Varian. 1986. On the private provision of public goods. *Journal of Public Economics* 29: 25–49.

Berk, R. 1974. A gaming approach to crowd behavior. *American Sociological Review* 39: 355–373.

Bernheim, B.D. 1986. On the voluntary and involuntary provision of public goods. *American Economic Review* 76: 789–793.

Blume, L.E. 1993. The statistical mechanics of strategic interaction. *Games and Economic Behavior* 5: 387–424.

Blume, L.E. 1995. The statistical mechanics of best-response strategy revision. *Games and Economic Behavior* 11: 111–145.

Blume, L.E. 1997. Population games. In *The economy as an evolving complex system II*, ed. W.B. Arthur, S.N. Durlauf, and D.A. Lane. Boulder: Westview.

Blume, L.E., and S.N. Durlauf. 2001. The interactions-based approach to socioeconomic behaviour. In *Social dynamics*, ed. S.N. Durlauf and H.P. Young. Cambridge, MA: MIT Press.

Blume, L.E., and S.N. Durlauf. 2003. Equilibrium concepts for social interaction models. *International Game Theory Review* 5: 193–209.

Brock, W.A., and S.N. Durlauf. 2001. Discrete choice with social interactions. *Review of Economic Studies* 68: 235–260.

Carlsson, H., and E. Van Damme. 1993. Global games and equilibrium selection. *Econometrica* 61: 989–1018.

Chamberlin, J. 1974. Provision of collective goods as a function of group size. *American Political Science Review* 68: 707–716.

Cornes, R. 1993. Dyke maintenance and other stories: Some neglected types of public goods. *Quarterly Journal of Economics* 108: 259–271.

Cornes, R., and T. Sandler. 1985. The simple analytics of pure public good provision. *Economica* 52: 103–116.

Cornes, R., and T. Sandler. 1996. *The theory of externalities, public goods and club goods*. 2nd ed. London: Cambridge University Press.

Cox, G.W. 1994. Strategic voting equilibria under the single nontransferable vote. *American Political Science Review* 88: 608–621.

Cox, G.W. 1997. *Making votes count*. Cambridge, UK: Cambridge University Press.

Diekmann, A. 1985. Volunteer's dilemma. *Journal of Conflict Resolution* 29: 605–610.

Foster, D., and H.P. Young. 1990. Stochastic evolutionary game dynamics. *Theoretical Population Biology* 38: 219–232.

Granovetter, M. 1978. Threshold models of collective behavior. *American Journal of Sociology* 83: 1420–1443.

Harsanyi, J.C., and R. Selten. 1988. *A general theory of equilibrium selection in games*. Cambridge, MA: MIT Press Classics.

Hirshleifer, J. 1983. From weakest-link to best-shot: The voluntary provision of public goods. *Public Choice* 41: 371–386.

Hirshleifer, J. 1985. From weakest-link to best-shot: Correction. *Public Choice* 46: 21–23.

Innes, R. 2006. A theory of consumer boycotts under symmetric information and imperfect competition. *Economic Journal* 116: 355–381.

Johnson, J.P. 2002. Open source software: Private provision of a public good. *Journal of Economics and Management Strategy* 11: 637–662.

Kandori, M., G.J. Mailath, and R. Rob. 1993. Learning, mutation and long-run equilibria in games. *Econometrica* 61: 29–56.

Kemp, M. 1984. A note on the theory of international transfers. *Economics Letters* 14: 259–262.

Lerner, J., and J. Tirole. 2001. The open source movement: Key research questions. *European Economic Review* 45: 819–826.

Lerner, J., and J. Tirole. 2002. Some simple economics of open source. *Journal of Industrial Economics* 50: 197–234.

Lipnowski, I., and S. Maital. 1983. Voluntary provision of a pure public good as the game of chicken. *Journal of Public Economics* 20: 381–386.

Marwell, G., and P.E. Oliver. 1993. *The critical mass in collective action: A micro-social theory*. Cambridge and New York: Cambridge University Press.

Marwell, G., P.E. Oliver, and R. Prahl. 1988. Social networks and collective action: A theory of the critical mass. III. *American Journal of Sociology* 94: 502–534.

McGuire, M. 1974. Group size, group homogeneity and aggregate provision of a pure public good under cournot behavior. *Public Choice* 18: 107–126.

McKelvey, R.D., and T.R. Palfrey. 1995. Quantal response equilibria for normal form games. *Games and Economic Behavior* 10: 6–38.
Monderer, D., and L.S. Shapley. 1996. Potential games. *Games and Economic Behavior* 14: 124–143.
Morris, S., and H.S. Shin. 2003. Global games: Theory and application. In *Advances in economics and econometrics: Theory and applications*, Eighth World Congress, ed. M. Dewatripont, L.P. Hansen and S.J. Turnovsky, vol. 1. London: Cambridge University Press.
Myatt, D.P. 2007. On the theory of strategic voting. *Review of Economic Studies* 74: 255–281.
Myerson, R., and R. Weber. 1993. A theory of voting equilibria. *American Political Science Review* 87: 102–114.
Oliver, P.E., and G. Marwell. 1988. The paradox of group size in collective action: A theory of critical mass. II. *American Sociological Review* 53: 1–8.
Oliver, P.E., G. Marwell, and R. Teixeira. 1985. A theory of the critical mass. I. Interdependence, group heterogeneity, and the production of collective action. *American Journal of Sociology* 91: 522–556.
Olson, M. 1965. *The logic of collective action: Public goods and the theory of groups*. Cambridge, MA: Harvard University Press.
Olson, M. 1987. Collective action. In *The new palgrave: A dictionary of economics*, ed. J. Eatwell, M. Milgate, and P. Newman, vol. 1. London: Macmillan.
Palfrey, T.R. 1989. A mathematical proof of Duverger's law. In *Models of strategic choice in politics*, ed. R. Ordeshook. Ann Arbor: University of Michigan Press.
Palfrey, T.R., and H. Rosenthal. 1984. Participation and the provision of discrete public goods: A strategic analysis. *Journal of Public Economics* 24: 171–193.
Raymond, E.S. 1998. The cathedral and the bazaar. *First Monday* 3, 1–20. Online. Available at http://www.firstmonday.dk/issues/issue3_3/raymond/. Accessed 19 April 2007.
Sandler, T. 2004. *Global collective action*. Cambridge: Cambridge University Press.
Schelling, T.C. 1973. Hockey helmets, concealed weapons, and daylight saving: A study of binary choices with externalities. *Journal of Conflict Resolution* 17: 381–428.
Schelling, T.C. 1978. *Micromotives and macrobehavior*. New York: Norton.
Stern, N. 2007. *The economics of climate change: The Stern review*. Cambridge: Cambridge University Press.
Warr, P. 1983. The private provision of a public good is independent of the distribution of income. *Economics Letters* 13: 207–211.
Young, D.J. 1982. Voluntary purchase of public goods. *Public Choice* 38: 73–85.
Young, H.P. 1993. The evolution of conventions. *Econometrica* 61: 57–84.
Young, H.P. 1998. *Individual strategy and social structure: An evolutionary theory of institutions*. Princeton: Princeton University Press.

Collective Agriculture

Peter Nollan

The socialist countries have generally modelled their rural institutions on those of the USSR in the 1930s. For the most part, means of production were owned by the so-called collective, farmwork was 'collectively' organized, and personal income 'collectively' distributed. At their peak, over one-third of the world's farmers worked under this system.

'Socialist' countries have favoured collectives for the following principal reasons.

Firstly, the leadership in most 'socialist' countries initially was afraid of an economically independent peasantry with ideas shaped by individualistic 'petty commodity production'. As Stalin put it: 'a great deal of work has to be done to remould the collective-farm peasant, to correct his individualistic mentality and to transform him into a real working member of a socialist society' (Stalin 1929, p. 469). Collectives were not intended as independent cooperatives: collectivization was party-led and collectives were subject to considerable external control (see e.g., Davies 1980; Volin 1970; Selden 1982; Unger 1984). Such a rationale is deeply undemocratic, especially given the peasants' numerical dominance in those countries (see, in particular, Cohen 1974, ch. 6).

Second, it was believed that state intervention through party-led collectives would improve rural economic performance (see e.g., Stalin 1929; General Office 1956). Collectives could raise savings and investment rates through reinvesting income and mobilizing 'surplus' labour for capital construction. Unfortunately, success in these respects can damage labour motivation by reducing current returns to collective labour. Collectives also could provide a vehicle for rapidly introducing new technology. However, this applies to bad as well as good technology – examples of the former are legion in 'socialist'

agriculture, including the various programmers in the Soviet Union associated with Lysenko (discussed in Volin 1970) and the ill-fated introduction of the double-wheeled, double-share plough, in China (Kuo 1972, ch. 12).

Third, party-led collectives were viewed as a means to attain high farm marketing rates and an outflow of farm sector savings to finance non-farm investment:

By transferring the disposal of agricultural output from individual peasants to government-supervised collective farm managements, collectivization destroys the basis for the peasants' resistance to the 'siphoning-off' of the economic surplus (Baran 1957, p. 268).

However, without, for example, adequate supplies of appropriately priced industrial commodities, forcibly raising the rate of farm sector marketings can reduce the growth rate of farm output and the future volume of farm marketings. Moreover, it has proved difficult to achieve a net farm savings outflow due, for example, to agriculture's need for industrial incentive goods and farm inputs (increased, insofar as inputs are inefficiently used and collectivization adversely affects livestock holdings, motive power and fertilizer supplies), and the state's inability to control private market prices (Ellman 1975; Ishikawa 1967).

Fourth, it was considered that collectives would prevent 'capitalist' polarization alongside farm modernization, with the majority of peasants becoming wage labourers (Stalin 1929; Mao 1955). Evidence from other developing countries contradicts Stalin and Mao's crude vision of rural class polarization (see, especially, Hayami and Kikuchi 1981). It indicates too that appropriate state policies (e.g. land reform, provision of education and credit, infrastructure construction, progressive taxation) can mitigate rural class inequalities. Class polarization is not the inevitable accompaniment of rural modernization, nor is collectivization the only way to resolve problems of rural class inequality (e.g. Hayami and Kikuchi 1981).

Fifth, Lenin, Stalin and Mao all believed that agriculture was characterized by lumpiness and economies of scale (Lenin 1899; Stalin 1929; Mao 1955). In many farm tasks, large scale is indeed an advantage, for example in research, processing, building and maintaining irrigation facilities. However, many modern farm inputs are divisible. Provided they are appropriately priced, credit is available and they have access to lumpy complementary inputs, all farm strata modernizing areas tend to acquire them (Hayami and Kikuchi 1981). Moreover, in large agricultural units labour supervision is a major problem (Bradley and Clark 1971). If a collective's members trust each other and are motivated to work hard for the group irrespective of relative income then labour supervision is not an issue. However, this is rarely the case (Morawetz 1983) and collective farm managers have had to devise payment systems to motivate farm workers. In certain farm tasks (notably harvesting) it is easy to pay labour according to its product, but for most farm tasks it is more difficult than in industry to devise payment systems that strongly motivate from the work often requires a flexible response from the worker which is difficult to anticipate in the payment system; the final produce takes a long time to produce, with different workers' contributions difficult to isolate; work is physically dispersed and production conditions vary greatly from one part of the production unit to another; the main task specializations are seasonal, and permanent minute sub-division of work into easily measurable segments is not generally possible. These problems have meant that under private agriculture, if labour is relatively abundant and capital relatively expensive, the normal outcome is for land to be rented out beyond a certain farm size, so that a relatively high output per acre can be attained through self-operating, self-motivated, rent-paying farmers, rather than cultivated with large numbers of hired workers. In collective farms, the attempt to supervise large numbers of farm workers has resulted in powerful managerial diseconomies of scale and reduced farm efficiency.

Collective agriculture has not performed well. Collective farms in the USSR in 1929–31 and in China in 1959–61 experienced massive institutionally caused declines in farm output, accompanied by demographic disasters (on the Soviet Union, see Volin 1970, ch. 10; on China, see Ashton et al. 1984). It is indeed, a terrible indictment of collective farming, that the worst famines

of the 20th century have occurred under that system. The USSRs long-term growth of farm output has required colossal capital outlays so that by the 1970s, the agricultural sector was absorbing over one quarter of Soviet new fixed investment (Carey 1976). From the mid-1950s to the later 1970s Chinese farm output per caput was stagnant: 'decollectivization' of agriculture in the early 1980s was accompanied by a huge in farm output (Nolan and Paine 1986).

The 'socialist' countries' poor agricultural performances is in part attributable to shortcomings in the supply of industrial goods (Smith 1981). Part is also due to extensive state intervention in collective farms. However, there are fundamental problems in principle even with relatively independent collective farms. Large units (whether state, collective or private) are necessary to undertake activities exhibiting lumpiness or economies of scale. However, for many farm tasks powerful managerial diseconomies of scale exist, and even given favourable policies in other respects, in most circumstances this would prove a barrier to good performance of collective farms.

See Also

- ▶ Agricultural Growth and Population Change
- ▶ Peasants

Bibliography

Ashton, B., et al. 1984. Famine in China, 1958–61. *Population and Development Review* 10(4): 613–645.

Baran, P. 1957. *The political economy of growth*. New York: Monthly Review Press.

Bradley, M.E., and M.G. Clark. 1972. Supervision and efficiency in socialized agriculture. *Soviet Studies* 23(3): 465–473.

Carey, D.W. 1976. Soviet agriculture: Recent performance and future plans. *In USCJEC* 1976.

Cohen, S.F. 1974. *Bukharin and the Bolshevik revolution*. London: Wildwood House.

Davies, R.W. 1980. *The socialist offensive: the collective action of Soviet agriculture, 1929–1930*. London: Macmillan.

Ellman, M. 1975. Did the agricultural surplus provide the resources for the increase in investment in the USSR during the First Five Year Plan? *Economic Journal* 85(4): 844–863.

General Office of the Central Committee of the Chinese Communist Party. 1956. *Socialist high tide in China's villages (Zhongguo nongcun de shehuizhuyi gaochao)*, vol 3 vols. Peking: People's Publishing House.

Hayami, Y., and M. Kikuchi. 1981. *Asian village economy at the crossroads*. Tokyo: University of Tokyo Press.

Ishikawa, S. 1967. Resource flow between agriculture and industry. *The Developing Economies* 5(1): 3–49.

Kuo, L.T.C. 1972. *The technical transformation of agriculture in communist China*. London: Praeger.

Lenin, V.I. 1899. *The development of capitalism in Russia*, 1964. Moscow: Progress Publishers.

Mao, Tsetung. 1955. On the co-operative transformation of agriculture. *Mao* 1977.

———. 1977. *Selected works of Mao Tsetung*, vol V. Peking: Foreign Languages Press.

Morawetz, D. 1983. The kibbutz as a model for developing countries. *Stewart* 1983.

Nolan, P., and S. Paine. 1986. Towards an appraisal of the impact of rural reform in China, 1978–85. *Cambridge Journal of Economics* 10(1): 83–99.

Selden, M. 1982. Co-operation and conflict: co-operative and collective formation in China's countryside. *Selden and Lippit* 1982.

Selden, M., and V. Lippit (ed). 1982. *The transition to socialism in China*. New York: M.E. Sharpe.

Smith, G.A.E. 1981. The industrial problems of Soviet agriculture. *Critique* 14: 41–65.

Stalin, J. 1929. Concerning questions of agrarian policy. In *Problems of Leninism*, ed. J. Stalin. Peking: Foreign Languages Press n.d.

Stewart, F. (ed). 1983. *Work, income and inequality*. London: Macmillan.

Unger, J. 1984. *Chen village*. Berkeley: University of California Press.

US Congress, Joint Economic Committee (USCJEC). 1976. *Soviet economy in a new perspective*. Washington, DC: US Government Printing Office.

Volin, L. 1970. *A century of Russian agriculture*. Cambridge, MA.: Harvard University Press.

Collective Bargaining

William Brown

Keywords

Arbitration; Bargaining; Collective bargaining; Collusion; Consultation; Employment contracts; Industrial democracy; Industrial relations; Negotiation; Trade unions; Wage determination

JEL Classifications
J5

Collective bargaining is a term applied to a variety of methods of regulating relationship between employers and their employees. Its distinctive feature is that it clearly acknowledges a role for trade unions. In contrast with, for example, autocratic paternalism or producer cooperatives, the employer who engages in collective bargaining accepts the right of independent representatives of employees, acting as a collectivity, to argue their point of view on matters that affect their interests. Pay and working conditions are the most common subjects of collective bargaining, but it can encompass any aspect of management.

The impact of collective bargaining upon management, and its effectiveness from the point of trade union members, vary enormously between different employment circumstances. They depend ultimately upon the collective strength that can be mobilized by employees within the legislative constraints laid down by the state. Collective bargaining is thus best seen as a political institution. It provides a means of bringing at least temporary reconciliation of divergent interests between employers and employees in circumstances in which each side can, to a greater or lesser extent, inflict damage on the other. It is, however, a political institution that is intimately linked with economic processes. The relative power of the bargaining partners owes much to their respective labour and product markets. At the same time the outcome of their bargaining has a major impact upon both the wages and the productivity of labour.

Theoretical Approaches

This view of collective bargaining as primarily a political rather than an economic institution is relatively recent. Beatrice Webb claimed, according to Marsh (1979), to have originated the expression in 1891 in her study *The Co-operative Movement of Great Britain*. She analysed it further with her husband Sidney Webb in *Industrial Democracy* (1897). Although they did not define it, they saw it as an alternative to individual bargaining, so that the employer, instead of making separate deals with isolated individuals, 'meets with a collective will and settles, in a single agreement, the principles on which, for the time being, all workmen of a particular group, or class, or grade, will be engaged'. They identified it as one of three methods used by trade unions to meet their objectives, the other two being to establish mutual assurance arrangements for their members and to press governments to enact favourable laws. For all the richness of the Webbs' analysis, collective bargaining remained for them essentially an economic institution, imposed upon the employer by a labour cartel whereby workers secured better terms of employment by controlling competition among themselves. A naive version of this view can be seen to underlie much formal analysis of collective bargaining by present-day labour economists.

For the next half century Marsh reports no substantial development of the concept apart from in Leiserson's *Constitutional Government in American Industries* (1922). Then in 1951 Chamberlain, in his book *Collective Bargaining,* argued that there were, in essence, three distinct theories. 'They are that collective bargaining is (1) a means of contracting for the sale of labour, (2) a form of industrial government, and (3) a method of management.' The first, 'marketing' theory was much the same as that of the Webbs. The second, 'governmental' aspect was concerned with the procedural needs of dispute resolution. The third 'managerial' theory referred to the way in which management and unions in practice combined 'in reaching decisions on matters in which both have vital interests'; unions through collective bargaining become not the usurpers of management functions but 'actually *de facto* managers'. At much the same time Harbison (1951) was stressing the very constructive social role that collective bargaining played in resolving industrial conflict and in pushing for the enhancement of the 'dignity, worth and freedom of individuals in their capacity as workers'.

This more complex view of collective bargaining has been refined by Dunlop (1967)

and Kochan (1980) in the United States, but probably the most influential discussion has been Flanders' attempt of 1968 to create a comprehensive theoretical analysis. He argued that the economic associations of the term 'collective bargaining' are misleading. The collective agreement commits no one to either buy or sell labour, but rather ensures that, when labour is bought or sold, the terms of the transaction will accord with the provisions of the agreement. Above all else, collective bargaining is a rule-making process covering many aspects of the employment relationship besides pay and conditions of work. The second characteristic feature of collective bargaining that Flanders stressed is that of the power relationship between the protagonists whose negotiations ('the diplomatic use of power') create the rules. Thus, while there are also technical rules and legal rules regulating work, what distinguishes the legitimacy of those that result from collective bargaining is their authorship. They are jointly determined by the accepted representatives of both employers and employees who consequently share responsibility for both the rules' contents and their observance.

Flanders' analysis has proved fertile in several respects. It has drawn attention to the extent to which collective bargaining is a positive management technique rather than just an impediment to effective management imposed by trade unions. As a result of this shift in emphasis, a major part of academic research into collective bargaining in the 1980s has explored managerial, as opposed to trade union, strategies, and has exposed the extent to which union behaviour is shaped by these management strategies. In addition, what could be seen as the Weberian undercurrent in Flanders' analysis has focused policymakers' attention upon the importance of procedural clarity in conflict resolution, and thereby upon the dangers of ambiguity in the legitimation of agreements. The most obvious example is provided by the influential central recommendation of the British Royal Commission on Trade Unions and Employers' Associations of 1968. The emphasis it placed upon employer initiated procedural reform, rather than legislative constraints on trade unions, owed much to the evidence that Flanders had submitted. Finally, by conceptualizing wages as part of a broader package of regulations and as embodying strongly normative principles, the theory opened the way to a more fruitful understanding of wage determination than is offered simply by the market models of orthodox theory.

Two crucial features of the employment relationship ensure that the process of collective bargaining is fundamentally unlike that of non-labour commercial bargains. They are its open-endedness and its continuity. The labour contract is open-ended because the recruitment of an employee does not ensure the performance of work; the employee has to be motivated, by whatever means, to perform to the required standard. In all but highly oppressive societies such motivational techniques tend to be varied and complex, differing not least in the extent to which they place emphasis upon levels of pay and upon employee participation. Since social comparisons (and especially very local ones) play an important part in the motivation and demotivation of labour, the bureaucratic standardization of terms of employment, which is generally a characteristic of collective agreements, often fits in well with management's preferred personnel techniques. In this way, properly conducted collective bargaining can provide a socially stable working environment which facilitates the employer's prime aim of eliciting labour productivity. In short, the conduct of the bargain affects the quality of the labour bargained over.

The second distinctive feature of the employment relationship is its continuity. Employer and employees are bound together, for better or worse, for an indeterminate duration. Additions to and departures from the workforce generally occur in a piecemeal way. A host of potentially contentious issues feature in the relationship, only a small minority in contention at any one time, and many affecting only a minority of the workforce. Thus a bargain over a particular issue, such as a pay grievance, cannot be evaluated in isolation, but as one fibre in a thick rope of regulations, with many largely implicit trade-offs with respect to other issues, past, present and future.

Characteristics

The definition of collective bargaining as the joint regulation of the employment relationship by employer and employee representatives is one that covers a broad range of processes. It is helpful to analyse these further. An initial distinction has to be made between negotiation and consultation. In a negotiation the discussions are characterized, first, by the awareness of each side of the possibility of one inflicting costs on the other in the absence of an acceptable outcome. Second, a negotiation has to result in some sort of agreement, however informal, to which the two sides are, at least for the time being, committed. Consultation, by contrast, is unaccompanied by either the threat of sanctions or the need to reach binding agreement. Actions taken by management in the light of consultation result from a reappraisal of the facts of the case; those taken after negotiation reflect a compromise which has taken into account the threat (or experience) of sanctions inflicted by either or both sides. Under most collective bargaining arrangements it is felt advisable by both sides to distinguish as far as is possible between negotiations and consultations, at any rate in formal procedures. It is, for example, now normal in large unionized workplaces in Britain to deal with them in specifically different committees, even though the membership of those committees may be much the same.

In practice the distinction is far from clear-cut. The blend of approaches adopted in a particular collective bargaining episode depends very much upon the issue in question and the relationship between the parties involved. In their study *A Behavioral Theory of Labor Negotiations* (1965), Walton and McKersie distinguished four classes of negotiation. First, there were 'distributive' bargains: zero-sum negotiations typified by annual wage bargains and characterized by very formal proceedings. Second were 'integrative' bargains: problem-solving discussions aiming at non-zero-sum gains for both sides and generally much more informal in procedure. Third, was 'attitudinal structuring', an almost didactic form of bargaining dialogue in which one side tries to alter the way in which their opponents perceive the problem and its context. Finally, 'intra-organizational' bargains were aimed at altering positions and attitudes, not on the other side, but within the negotiator's own side.

An important influence upon the way in which bargaining is conducted is the personal 'bargaining relationship' between the two individuals who have to take the lead in representing the two sides. This is a term given to the level of trust and facility of communication that exists between them. However acrimonious the collective dispute over which they are bargaining, the better the bargaining relationship between the individual negotiators, the more efficiently they will be able to assess each other's relative power position and the better the chance of the dispute being settled without recourse to expensive sanctions. In a mature bargaining relationship it is common for the negotiators to protect each other from their own sides by, for example, avoiding the humiliation of a bargaining opponent by helping him to gloss over the magnitude of a defeat and by manipulating public statements from one's own side so as to help in his intra-organizational bargaining with his own.

It is normal to draw a clear distinction between the substantive and procedural aspects of collective bargaining. A substantive agreement sets out the actual pay levels, working conditions, or whatever that have been agreed and will be worked to. A procedural agreement defines the way in which such substantive terms might be altered, added to, or interpreted. An effective procedure for negotiation or grievance settlement will state which agents on each side are entitled to be involved in negotiations, in what sequence different sets of negotiators are entitled to consider the matter, what their precedence is, and possibly also matters such as rights of appeal, time constraints, ratification methods and the form of the substantive outcome.

This distinction is particularly obvious in countries whose labour laws cause collective agreements to be tested in the courts; the substantive agreements tend to be written, detailed, formal, and established for specified duration. There

are other countries where employer preference, or legal opportunity, makes it unusual for the bargaining opponents to use legal sanctions against each other. In these circumstances the great bulk of substantive regulation may be unwritten and in the form of verbal agreements, custom, and tacit understandings. Because of this a greater emphasis is placed upon the rectitude of the procedural agreements (which may still be very informal) whereby this amorphous body of substantive rules is interpreted and altered, not through comprehensive periodic negotiations, but by a constant incremental process of piecemeal adjustment. Although the United States might be described as exemplifying the legalistic extreme, and Great Britain the 'voluntaristic', most bargaining arrangements have elements of each, with the degree of legalism and formality varying by issue and industry, as well as by country.

Bargaining Structure

The structure of bargaining in a country, industry, or enterprise, refers to several different characteristics of collective bargaining. The two most important are the 'bargaining units' and 'bargaining levels' employed. A bargaining unit is a group of employees covered by a particular agreement. Within this basic territory of industrial government there is a coherence of terms of employment, procedures, and trade union representation that is not necessarily to be found between different bargaining units. The level of bargaining refers to the role played by the principal negotiators within their organizations; whether, for example, the employer representative responsible is a factory manager, a company director, or an employers' association representative.

These two characteristics are involved in the single most important decision in the shaping of any bargaining structure which is whether the employers confront the unions singly or in alliance. Single-employer bargaining, resulting in agreements at company-level or lower, is the majority practice in the United States and Japan and now in Britain. Multi-employer bargaining, in which associations of employers conclude industrywide agreements, remains the most important form in most of Continental Europe. In practice there is often some employer collusion in industries where single-employer bargaining dominates, and there is usually room for individual employer discretion in industries with strong employers' associations, but the distinction remains one of fundamental economic, political, and managerial significance.

Two other defining characteristics of bargaining structure are its 'form' and 'scope'. The first refers to the extent to which proceedings and agreements are formalized and codified. As already mentioned, this depends in part upon the labour legislation of the country. The second matter, scope, refers to the range of issues covered by collective bargaining. At its narrowest it may include no more than pay and hours, while elsewhere it may take in issues as diverse as training policy, investment decisions and child-care facilities.

The most comprehensive theory seeking to explain industrial and national differences in bargaining structure is to be found in Clegg's *Trade Unionism under Collective Bargaining* (1976). This sees the strategy adopted by employers as the main determinant of bargaining structure, although changes in strategy may be slow to take effect. The legislative framework of a country is also of crucial importance. It defines the limits of rights to strike, the status of the employment contract, any guarantees of security for trade unions, and the legally responsible agents on each side.

Most countries acquired their principal labour legislation at some historic period of crisis – war, defeat, depression, or extreme industrial unrest – and the institutional arrangements that developed from that have become consolidated in subsequent, more peaceful times. This helps to account for the very great variations in collective bargaining practice to be found in different countries; they often owe their origin to a distant panic measure based upon a fashionable idea (such as, for example, compulsory arbitration in Australia or compulsory conciliation in Canada) to which employers and unions have adjusted so firmly that radical reformation is all but impossible. A recurring experience around the world is of

legislatures finding extreme difficulty in reforming collective bargaining, other than in times of extreme crisis, because of the essential privacy of the bargaining relationship between employers and union.

Most industrialized countries publicly assert a commitment to collective bargaining as a necessary part of a democratic society, and for most it is the normal means of conducting industrial relations in the public sector. Convention 84 (1947) of the International Labour Organization asserts that 'all practical measures shall be taken to assure to trade unions which are representative of the workers concerned the right to conclude collective agreements with employers and employers' associations'. In practice the freedom of collective bargaining in both public and private sectors varies substantially between countries and over time.

No discussion of collective bargaining would be complete without a mention of the debate concerning its relationship with industrial democracy.

One view is that, because collective bargaining is essentially concerned with compromise, trade unions are sucked into collaborating with capitalism and thereby denied the opportunity of uniting the working class in overthrowing existing employers and then instituting true industrial democracy through workers' control. Opposing this is a view that deplores the fact that collective bargaining institutionalizes the opposition of capital and labour: them and us. It considers that the best form of industrial democracy is to be found where workers are brought to perceive an ultimate identity of interest with employers. Between these positions is that most clearly expressed by Clegg in *A New Approach to Industrial Democracy* (1960). This argues that there can never be complete identity of interest between employer and employee, and also that if employee representatives are given managerial responsibilities they will be forced to behave very similarly to the employers they have replaced. Consequently the role of the trade union is best seen as one of constant opposition, acting to modify management actions in the light of members' interests insofar as their organized power permits. Far from undermining the common interests of capital and labour, collective bargaining permits the joint regulation of aspects of employment which would otherwise generate greater disharmony and division.

See Also

► Industrial Relations

Bibliography

Chamberlain, N.W. 1951. *Collective bargaining*. New York: McGraw-Hill.
Clegg, H.A. 1960. *A new approach to industrial democracy*. Oxford: Blackwell.
Clegg, H.A. 1976. *Trade unionism under collective bargaining*. Oxford: Blackwell.
Dunlop, J.T. 1967. The social utility of collective bargaining. In *Challenges to collective bargaining*, ed. L. Ulman. New York: Prentice-Hall.
Flanders, A. 1968. Collective bargaining: A theoretical analysis. *British Journal of Industrial Relations* 6 (1): 1–26. Reprinted in Flanders, A. 1975. *Management and unions*. London: Faber & Faber.
Harbison, F.H. 1951. *Goals and strategies in collective bargaining*. New York: Harper.
Kochan, T.A. 1980. *Collective bargaining and industrial relations*. Homewood: Irwin.
Leiserson, W.M. 1922. Constitutional government in American industries. *American Economic Review* 12 (Suppl): 56–79.
Marsh, A. 1979. *Concise encyclopedia of industrial relations*. Farnborough: Gower.
Walton, R.E., and R.B. McKersie. 1965. *A behavioral theory of labor negotiations*. New York: McGraw-Hill.
Webb, S., and B. Webb. 1897. *Industrial democracy*. London: Longmans Green.

Collective Choice Experiments

Rick K. Wilson

Abstract

Collective choice experiments examine voting mechanisms that aggregate individual preferences. Two general topics have received the most attention. The first pertains to agents deciding on a single collective outcome or

policy. The second topic covers election mechanisms that govern candidates and voters.

Keywords
Agenda control; Arrows; Theorem; Collective choice experiments; Electoral mechanisms; Median voter; Social choice; Spatial committee experiments

JEL Classifications
C9

Duncan Black (1948) and Kenneth Arrow (1963) raised the key question of collective choice: if people have different preferences for policy outcomes are there general mechanisms that can (always) aggregate those preferences in consistent and coherent ways? The answer is 'no'. Starting from simple premises involving individual transitivity, aggregate Pareto optimality and non-dictatorship there is no collective choice mechanism that yields a socially transitive outcome. Such a finding is startling given the confidence placed in democratic institutions that rely on voting mechanisms to choose a single outcome from many possible outcomes.

Experimentalists have thoroughly explored different institutions that can be used to aggregate preferences. Political economists who straddle both economics and political science have carried out much of this work. Their concern is with situations where actors who have opposed interests have to settle on a single outcome and with the properties of the institution used to produce an outcome. This article first turns to the institutional mechanisms by which individuals settle on a collective outcome. The second topic turns to electoral mechanisms used in representative democracies.

Spatial Committee Experiments

In the late 1960s theoretical papers by Davis and Hinich (1966) and Plott (1967) described a social choice environment for spatial committees. Those committees consist of a well-defined multidimensional policy space, with actors holding fixed preferences over the dimensions, and policies represented as points in the space. Using rules that mimic many parliamentary systems, these theoretical papers demonstrate that a Condorcet winner (a policy that can defeat all others under pairwise voting) exists only under rare distributions of voters' preferences. Plott (1967) establishes the conditions under which a Condorcet winner will exist and he makes the connection between this and a Nash equilibrium of a spatial committee game. Like others, he concludes that an equilibrium is rare in multidimensional spatial committee games.

Early spatial committee experiments by Berl et al. (1976) and Fiorina and Plott (1978) provide evidence that when a Condorcet winner exists, subjects choose it or outcomes that are close to it. In games where there is no such equilibrium (which is the most common case), subjects select outcomes that scatter in the policy space. These initial empirical findings, coupled with experiments by Laing and Olmsted (1978) and McKelvey et al. (1978), defined the standard for conducting spatial committee experiments. Subsequent experiments have adopted almost identical procedures.

The standard experimental design introduces a two-dimensional policy space. The orthogonal dimensions are arbitrary (X and Y in most settings) and typically range from zero to 200 or more units. Every point in the space characterizes a policy. Preferences over outcomes are induced by assigning each subject a payoff function mapping earnings in dollars to each point in the space. While many payoff functions have been tested, most experimenters have settled on a quadratic loss function, with monetary payoffs decreasing as a function of distance from a subject's ideal point. Usually five subjects are assigned different ideal points in the space, and it is the arrangement of these ideal points that allows the experimenter to manipulate, whether a Condorcet winner exists or not. Subjects are given an initial status quo and then allowed to introduce amendments. Voting takes place following an amendment, with the winner becoming (or remaining) the new status quo. Amending takes place in between votes. A motion to adjourn, passed under a voting rule, constitutes the stopping rule for the committee decision. This serves as the standard institution

for subsequent spatial committee experiments. Changing these basic institutional rules became the way to test theories of collective choice.

Experimental results in the absence of equilibrium are both frustrating and profitable. Frustration arises over the fact that committee choices tend to be clustered in similar regions of the policy space. While there appears to be some pattern to the outcomes, the process by which these outcomes arise has not been fully characterized (but see the attempt by Bianco et al. 2006). Profitably, these empirical results led theorists and experimenters to add agenda control to the structure of the game. This led to a distinction between preference-induced and structure-induced equilibrium. For example, Plott and Levine (1978) showed the effectiveness of agenda control both in the laboratory and in a natural setting. Awarding agenda power created a structure-induced equilibrium and laboratory subjects converged to it. Recent experimental work by Frechette et al. (2003) illustrates that the equilibrium favours agenda setters.

Theoretical work by Buchanan and Tullock (1962) led experimentalists to examine whether changing the proportion of actors needed to pass a policy had any effect. Experiments by Laing and Slotznick (1983) showed that moving from simple majority rule (50 per cent plus 1) to supermajority majority rule (67 per cent) resulted in many equilibria and that subjects chose them. Schofield (1985), among others, provided the theoretical basis for when an equilibrium exists as a function of the dimensionality of the policy space, the voting rule and the distribution of voters' preferences. These theoretical findings spurred experimentalists to examine other changes to the standard committee experiment. For example Wilson and Herzberg (1987) theoretically predicted and experimentally demonstrated that when a single player holds veto power, that player's ideal point is the equilibrium. Haney et al. (1992) empirically show committee choices converging to equilibrium when a weighted voting rule is used. Such a rule requires that a single player always be included in a coalition. These results are representative of the kind of work that has dominated the experimental spatial committee agenda.

Experiments on spatial committees have added to a clearer understanding of institutional mechanisms. Experimental results demonstrate that changing who has the power to set the agenda, how the agenda is built, how many votes are needed and whether players enjoy veto powers, matters.

C

Electoral Mechanisms

A second area of interest for collective choice experimentalists is with electoral mechanisms. Three broad directions have been taken that treat different aspects of representative democracies. The first is concerned with candidate behaviour. At the heart of this research is the question of whether candidate positions will converge to equilibrium when it exists. The second direction is concerned with voter behaviour, particularly how voters behave when they have little information about candidate positions. The final direction deals with the way in which electoral rules determine the likelihood that 'types' of candidates are elected, where types usually refer to racial and ethnic minority candidates.

The initial experimental work on candidate behaviour focused on candidates who cared only about winning and varied the information conditions that the candidates have about the preferences of voters. Most experiments use a unidimensional policy space that guarantees an equilibrium. This equilibrium is defined by the policy preference of the median voter. In the experiments elections are sequential, with two candidates announcing positions in the policy space and voters choosing between the candidates. Voters are assigned ideal points in the policy space, the winning candidate is required to implement the announced policy and voters are paid an amount that decreases with the distance of the winning position from their ideal point. Candidates are paid only if they win. Once the election is over another election is held with candidates free to change their previously announced policy. Not surprisingly, all candidates quickly adopt the position of the median voter when they are fully informed about voter preferences. Under incomplete information about voters,

candidates also converge to the median voter's position, by responding to feedback about the vote share accruing to different policy positions, as in McKelvey and Ordeshook (1985). If candidates have policy preferences whereby their earnings depend not only on winning but also on implementing a policy close to their own preferred position, then the median voter result no longer holds (see the experimental results by Morton 1993).

When voters are uninformed about candidate positions, are they able to cast accurate ballots? With minimal information, such as biased endorsements or polls, subjects do very well at inferring candidate positions. Lupia and McCubbins (1998) and Morton and Williams (2001) consider various aspects of voter information and show that voters are able to quickly determine the positions of candidates and cast their vote accordingly.

Finally, several experiments have focused on differing electoral mechanisms and what they mean for the type of candidates that gain election. For example Gerber et al. (1998) compare two voting mechanisms in an experiment to test whether one or the other disadvantages a racial or ethnic minority candidate. A form of cumulative voting (in which voters can cast more than a single vote) leads to more minority candidates being elected. This should be no surprise to collective choice theorists who have long noted that different electoral mechanisms lead to predictable variation in outcomes. Cox (1997) offers an extended discussion of such mechanisms.

What We Know

Collective choice experiments provide several insights. First, when a Nash equilibrium of the underlying game exists it is a strong predictor of the outcome of the experiment. The second finding is that when there is no Nash equilibrium for the underlying game, subjects choose outcomes that cluster in predictable areas of the policy space, but the process by which that occurs is not settled. At the same time, experimentalists have implemented institutional mechanisms altering such games, thereby producing an equilibrium that subjects choose. Often those institutional changes benefit one actor (for example, by assigning agenda control to a particular player). A third finding is that incomplete information does not prevent convergence to equilibrium for either candidate platform choice or voter behaviour. The fourth finding returns to Arrow's original insight: voting mechanisms can be manipulated to achieve predictable, but very different, outcomes. It all depends on the mechanism that is implemented.

See Also

- ▶ Arrow's Theorem
- ▶ Experimental Economics
- ▶ Political Institutions, Economic Approaches to
- ▶ Social Choice
- ▶ Strategic Voting
- ▶ Voting Paradoxes

Bibliography

Arrow, K.J. 1963. *Social choice and individual values*. New Haven: Yale University Press.

Berl, J.E., R.D. McKelvey, P.C. Ordeshook, and M. D. Winer. 1976. An experimental test of the core in a simple n-person cooperative nonsidepayment game. *Journal of Conflict Resolution* 20: 453–476.

Bianco, W.T., M.S. Lynch, G.J. Miller, and I. Sened. 2006. 'A theory waiting to be discovered and used': A reanalysis of canonical experiments on majority rule decision making. *Journal of Politics* 68: 837–850.

Black, D. 1948. On the rationale of group decision making. *Journal of Political Economy* 56: 22–34.

Buchanan, J., and G. Tullock. 1962. *Calculus of consent*. Ann Arbor: University of Michigan Press.

Cox, G.W. 1997. *Making votes count: Strategic coordination in the world's electoral systems*. New York: Cambridge University Press.

Davis, O.A., and M.J. Hinich. 1966. A mathematical model of policy formation in a democratic society. In *Mathematical applications in political science*, ed. J.L. Bernd. Dallas, TX: Southern Methodist University.

Fiorina, M.P., and C.R. Plott. 1978. Committee decisions under majority rule: an experimental study. *American Political Science Review* 72: 575–598.

Frechette, G., J.H. Kagel, and J.H. Lehrer. 2003. Bargaining in legislatures: An experimental investigation of open versus closed amendment rules. *American Political Science Review* 97: 221–232.

Gerber, E.R., R.B. Morton, and T.A. Rietz. 1998. Minority representation in multimember districts. *American Political Science Review* 92: 127–144.
Haney, P., R. Herzberg, and R.K. Wilson. 1992. Advice and consent: Unitary actors, advisory models and experimental tests. *Journal of Conflict Resolution* 36: 603–633.
Laing, J.D., and S. Olmsted. 1978. An experimental and game theoretic study of committees. In *Game Theory and Political Science*, ed. P.C. Ordeshook. New York: New York University Press.
Laing, J.D., and B. Slotznick. 1983. Winners, blockers, and the status quo: Simple collective decision games and the core. *Public Choice* 40: 263–279.
Lupia, A., and M.D. McCubbins. 1998. *The democratic dilemma: Can citizens learn what they need to know?* Cambridge: Cambridge University Press.
McKelvey, R.D., and P.C. Ordeshook. 1985. Sequential elections with limited information. *American Journal of Political Science* 29: 480–512.
McKelvey, R.D., P.C. Ordeshook, and M.D. Winer. 1978. The competitive solution for n-person games without transferable utility with an application to competitive games. *American Political Science Review* 72: 599–615.
Morton, R.B. 1993. Incomplete information and ideological explanations of platform divergence. *American Political Science Review* 87: 382–392.
Morton, R.B., and K.C. Williams. 2001. *Learning by voting: Sequential choices in presidential primaries and other elections*. Ann Arbor, MI: University of Michigan Press.
Plott, C.R. 1967. A notion of equilibrium and its possibility under majority rule. *American Economic Review* 57: 787–806.
Plott, C.R., and M.E. Levine. 1978. A model of agenda influence on committee decisions. *American Economic Review* 68: 146–160.
Schofield, N. 1985. *Social Choice and Democracy*. Heidelberg: Springer.
Wilson, R.K., and R.Q. Herzberg. 1987. Negative decision powers and institutional equilibrium: Experiments on blocking coalitions. *The Western Political Quarterly* 40: 593–609.

Collective Models of the Household

Olivier Donni

Abstract

Collective models of the household are based on two fundamental assumptions: (*a*) each agent is characterized by specific preferences and (*b*) the decision process results in Pareto-efficient outcomes. The main results of the theory of collective models then refer to the empirical issue of deriving testable restrictions on household behaviour and recovering from this some information on the structural model that can be used to carry out welfare comparisons at the individual level.

Keywords

Collective models of the household; Exclusive goods; Household behaviour; Indirect utility function; Pareto efficiency

JEL Classifications

D11

Until recently 'unitary' models, which assume that household members act as if they maximize a unique utility function under a budget constraint, were largely predominant in the literature on household behaviour. There is increasing agreement, however, that economists cannot ignore the fact that most households are composed of several individuals who take part in the decision process. Consequently, the 'collective' models, which postulate that (*a*) each household member has specific, generally different preferences and (*b*) the decision process results in Pareto-efficient outcomes, have attracted considerable attention from the profession during recent years.

To examine the properties of collective models, let us consider a household consisting of two persons, *A* and *B*, who make decisions about consumption. These persons are characterized by well-behaved utility functions of the form: $u_i(\mathbf{x}_A, \mathbf{x}_B, \mathbf{X})$, where $\mathbf{x}_i$ denotes a vector of private goods consumed by member i and $\mathbf{X}$ a vector of public goods ($i = A, B$). This specification of preferences is very general; it allows for altruism but also for externalities or any other preference interaction. We denote the vector of prices for private goods by $\mathbf{p}$, the vector of prices for public goods by $\mathbf{P}$ and the household total expenditure by y. Finally, we suppose that there exists a vector of distribution factors, that is, a set of exogenous variables which influence the intra-household

allocation of resources without affecting preferences or the budget constraint. Examples are given by the respective contribution of each member to the exogenous household income, the state of the marriage market or divorce legislation. These variables, which are often assigned a crucial role in the derivation of the results, are denoted by **s**.

To simplify notation, let $\pi' = (\mathbf{p}', \mathbf{P}')$ be the vector of prices. Then, efficiency essentially means that household behaviour can be described by the maximization of a utilitarian social welfare function, that is,

$$\max_{\mathbf{x}_A, \mathbf{x}_B, \mathbf{X}} \mu(\pi, y, \mathbf{s}) u_A(\mathbf{x}_A, \mathbf{x}_B, \mathbf{X}) + (1 - \mu(\pi, y, \mathbf{s})) u_B(\mathbf{x}_A, \mathbf{x}_B, \mathbf{X}) \quad (1)$$

subject to $\mathbf{p}'(\mathbf{x}_A + \mathbf{x}_B) + \mathbf{P}'\mathbf{X} = y$. In this programme, the function μ determines the location of the household equilibrium along the Pareto frontier. If $\mu = 1$, then the household behaves as though member A always gets her way whereas, if $\mu = 0$, it is as if member B is the effective dictator. We denote the solutions to Eq. (1) by $\mathbf{x}_A(\pi, y, \mathbf{s})$, $\mathbf{x}_B(\pi, y, \mathbf{s})$ and $\mathbf{X}(\pi, y, \mathbf{s})$.

Characterization

The first objective of the theory of collective models is to investigate the properties of the household demands derived from Eq. (1). These properties can either be tested statistically or be imposed a priori for simplifying the estimation task. From this perspective, one crucial point is that individual demands for private goods, $\mathbf{x}_A$ and $\mathbf{x}_B$, are generally unobservable by the outside econometrician; demands for these goods are observed only at the household level, $\mathbf{x} = \mathbf{x}_A + \mathbf{x}_B$ To be useful, the restrictions derived from the collective setting have thus to characterize household demands, **x** or **X**, instead of individual demands, $\mathbf{x}_A$ and $\mathbf{x}_B$.

Let $\boldsymbol{\xi} = (\mathbf{x}', \mathrm{X}')$ be the vector of household demands. We define the Pseudo–Slutsky matrix as follows:

$$\mathbf{S} = \frac{\partial \boldsymbol{\xi}}{\partial \pi'} + \frac{\partial \boldsymbol{\xi}}{\partial y} \boldsymbol{\xi}'$$

There exist at least three different sets of testable restrictions that characterize household behaviour.

SR1 Condition

Browning and Chiappori (1998) and Chiappori and Ekeland (2006) show that household demands compatible with Eq. (1) have to satisfy the following condition:

$$\mathbf{S} = \sum + \mathbf{R}_1,$$

where Σ is a symmetric, semi-definite matrix and $\mathbf{R}_1$ is a rank one matrix. The interpretation is the following. For any given pair of utility functions, (a) the budget constraint determines the Pareto frontier as a function of π and y, and (*b*) the value of μ determines the location of the household equilibrium on this frontier. Consequently, a change in π implies a shift of the Pareto frontier. The latter entails the modification of household demands described by Σ. However, the value of μ varies as well, hence the location of the equilibrium moves along the Pareto frontier. Since the frontier is of dimension one, this effect is very restricted and defined by $\mathbf{R}_1$.

Proportionality Condition

The particular structure of Eq. (1) leads to further restrictions on behaviour. To make things simple, let us suppose that the vector of distribution factors is twodimensional: $\mathbf{s} = (s_1, s_2)$. Then, Bourguignon et al. (1993) demonstrate the following result:

$$\frac{\partial \boldsymbol{\xi}}{\partial s_1} = \theta \frac{\partial \boldsymbol{\xi}}{\partial s_2},$$

where θ is a scalar. Thus, the response to different distribution factors is co-linear. The interpretation is that distribution factors can only change the location of the outcome on the frontier (through function μ), and the latter is of dimension one.

Specific Conditions

The econometrician is often inclined to put more structure on preferences. For example, let us suppose that agents have utility functions of the form: $u_i(\mathbf{x}_i, \mathbf{X})$. In that case, we say that agents are

'egoistic' in the sense that the utility does not depend on the partner's consumption. This assumption implies, in particular, that the decision process can be decentralized. In a first step, household members agree on the level of public goods as well as on a particular distribution of the residual expenditure between them. In a second step, they maximize their utility function, taking into account the level of public goods and their own budget constraint. It means, formally, that there exists a pair of functions ($\rho_A(\mathbf{p}, \mathbf{X}, y^*, \mathbf{s})$, $\rho_B(\mathbf{p}, \mathbf{X}, y^*, \mathbf{s})$), satisfying $\rho_A + \rho_B = y^*$ where $y^* = y - \mathbf{P}'\mathbf{X}$, such that the demand for private goods by member i is the solution to

$$\max_{\mathbf{x}_i} u_i(\mathbf{x}_i, \mathbf{X}) \text{ subject to } \mathbf{p}'\mathbf{x}_i = \rho_i.$$

Hence, household demands for private goods, conditionally on the demands for public goods, can be written as:

$$\mathbf{x} = \mathbf{x}_A(\mathbf{p}, \mathbf{X}, \rho(\mathbf{p}, \mathbf{X}, y^*, \mathbf{s})) + \mathbf{x}_B(\mathbf{p}, \mathbf{X}, y^* - \rho(\mathbf{p}, \mathbf{X}, y^*, \mathbf{s})),$$

where $\rho = \rho_A$ and $y^* - \rho = \rho_B$. This structure generates strong testable restrictions because the same function $\rho(\mathbf{p}, \mathbf{X}, y^*, \mathbf{s})$ enters each demand for private goods. Bourguignon, Browning and Chiappori (1995) explicitly derive these restrictions under the form of partial differential equations, whereas Donni (2004) shows that the demands for public goods have a particular but different structure, which implies testable restrictions as well.

Welfare Analyses – Identification

One of the main sources of interest in collective models is to provide the theoretical background for performing welfare comparisons at the individual level. The key concept in that case is what Chiappori (1992) calls the 'collective' indirect utility function. Let us suppose again that agents are egoistic. If so, the collective indirect utility function is defined as follows:

$$v_i(\pi, y, \mathbf{s}) = u_i(\mathbf{x}_i(\pi, y, \mathbf{s}), \mathbf{X}(\pi, y, \mathbf{s})). \quad (2)$$

This expression describes the level of welfare that member i attains in the household when he or she faces the price-income bundle(π, y) and a set of distribution factors $\mathbf{s}$. This representation of utility differs from the 'unitary' indirect utility function in that it implicitly includes the sharing function, and hence an outcome of the collective decision process. However, the knowledge of Eq. (2) is usually sufficient to evaluate the impact of economic policies on individual welfare.

In general, if agents are egoistic, the collective indirect utility functions can be retrieved. Nonetheless, the econometrician must observe the demand for some specific goods, referred to as 'exclusive', which benefit only one person in the household. More precisely, we say that good $X(x)$ is exclusively consumed by member i if $\partial u_j/\partial X = 0 (\partial u_j/\partial x_j = 0)$ for $j \neq i$. The intuition is that the household demand for 'exclusive' goods can be used as an indicator of the distribution of bargaining power within the household. Donni (2006) considers the case of purely private consumption ($\mathbf{X} = 0$) and shows that, if there is a single exclusive good, the collective indirect utility functions can be identified up to composition by an increasing transformation. Similarly, Chiappori and Ekeland (2003) consider the opposite case of purely public consumption ($\mathbf{x} = 0$) and show that, if there are two exclusive goods (one for each member), the identification is still possible. However, the general case with both private and public consumption has not been completely treated until now; see Blundell, Chiappori and Meghir (2005) for a first investigation.

Bibliographical Note

The main idea of collective models can be traced back to Leuthold (1968), who estimates a model of household labour supply based on non-cooperative game theory, where the individual is the basic decision-maker. However, this model differs from collective models in that the underlying decision process does not result in efficient outcomes. It actually belongs to the family of 'strategic' models (which are sometimes

referred to as 'collective' models in a broad sense). Nevertheless, a significant advance towards the development of collective models is made by Manser and Brown (1980) and McElroy and Horney (1981) at the beginning of the 1980s. These authors study the properties of models based on bargaining theory, which implies Pareto-efficiency. In that case, the location along the Pareto frontier is determined by the Nash (or Kalai–Smorodinsky) solution. However, the first formal investigation of a model based on the sole efficiency assumption is due to Chiappori (1988, 1992) in the context of labour supply decisions. This model is not explicitly examined in this article because it can be seen as a particular case of the model of consumption. Note, however, that Apps and Rees (1997), Chiappori (1997), Donni (2003), and Fong and Zhang (2001) present theoretical extensions of Chiappori's initial model, whereas Chiappori, Fortin and Lacroix (2002) exhibit empirical results. Finally, we must mention that several authors have generalized collective models to inter-temporal decisions and uncertain environment. One of the most representative examples of these studies is given by Mazzocco (2005).

See Also

- Family Decision Making
- Gender Roles and Division of Labour
- Household Production and Public Goods
- Household Surveys
- Individualism Versus Holism
- Integrability of Demand
- Intrahousehold Welfare
- Labour Supply
- Rotten Kid Theorem

Bibliography

Apps, P., and R. Rees. 1997. Collective labor supply and household production. *Journal of Political Economy* 105: 178–190.

Blundell, R., P.-A. Chiappori, and C. Meghir. 2005. Collective labor supply with children. *Journal of Political Economy* 113: 1277–1306.

Bourguignon, F., Browning, M., and P.-A. Chiappori. 1995. The collective approach to household behaviour. Working Paper. Paris: DELTA.

Bourguignon, F., M. Browning, P.-A. Chiappori, and V. Lechene. 1993. Intrahousehold allocation of consumption: A model and some evidence from French data. *Annales d'économie et de statistique* 29: 137–156.

Browning, M., and P.-A. Chiappori. 1998. Efficient intrahousehold allocations: A general characterization and empirical tests. *Econometrica* 66: 1241–1278.

Chiappori, P.-A. 1988. Rational household labor supply. *Econometrica* 56: 63–89.

Chiappori, P.-A. 1992. Collective labor supply and welfare. *Journal of Political Economy* 100: 437–467.

Chiappori, P.-A. 1997. Introducing household production in collective models of labor supply. *Journal of Political Economy* 105: 191–209.

Chiappori, P.-A., and I. Ekeland. 2003. The micro economics of group behavior: Identification. Working paper. Chicago: University of Chicago.

Chiappori, P.-A., and I. Ekeland 2006. Characterizing group behavior. *Journal of Economic Theory* (forthcoming).

Chiappori, P.-A., B. Fortin, and G. Lacroix. 2002. Marriage market, divorce legislation, and household labor supply. *Journal of Political Economy* 110: 37–72.

Donni, O. 2003. Collective household labor supply: Non-participation and income taxation. *Journal of Public Economics* 87: 1179–1198.

Donni, O. 2004. The intrahousehold allocation of private and public consumption:Theory and some evidence from U.S. data. Working paper. Cergy-Pontoise: University of Cergy-Pontoise.

Donni, O. 2006. Collective consumption and welfare. *Canadian Journal of Economics* 39: 124–144.

Fong, Y., and J. Zhang. 2001. The identification of unobservable independent and spousal leisure. *Journal of Political Economy* 109: 191–202.

Leuthold, J. 1968. An empirical study of formula transfers and the work decision of the poor. *Journal of Human Resources* 1: 312–323.

Manser, M., and M. Brown. 1980. Marriage and household decision making: A bargaining analysis. *International Economic Review* 21: 31–44.

Mazzocco, M. 2005. Household intertemporal behavior: a collective characterization and empirical tests. Working paper. Madison: University of Wisconsin.

McElroy, M., and M. Horney. 1981. Nash bargained household decisions. *International Economic Review* 22: 333–349.

Collective Rationality

Lu Hong

Abstract

This article reviews the concepts of individual rationality and collective rationality as they appear in the economics literature. In particular, the existing literature on social choice and aggregate demand points to a fundamental disconnect between these two notions of rationality. A possible reconciliation of this disconnect is suggested.

Keywords

Aggregate demand; Arrow's impossibility theorem; Collective choice; Collective rationality; Debreu–Mantel–Sonnenschein theorem; Gibbard–Satterthwaite impossibility theorem; Individual rationality; Prisoner's dilemma; Rational choice; Sen, A.; Social choice; Social welfare function; Strategic behaviour

JEL Classifications

D02; D71; D72; D81; D82

Since ancient times, men have argued that choice should be governed by 'desire and reasoning directed to some end' (Sen 1995). Much modern economic theory is based on this rational choice principle paradigm. In an individual choice problem, the individual is assumed to have a preference ordering on the set of alternatives. The individual choice is rational if, for any given decision situation, the choice made is always the best among all feasible alternatives according to the preference ordering. In a collective choice problem, be it that of a society or a committee, the definition of this rational choice principle becomes problematic. As there is presumably a huge disparity among the desires and ends of the individuals within the collective, by whose desire and whose end should the collective choice be governed? Is it reasonable to expect the collective choice to be guided by a preference ordering? If so, how should it reflect individual preferences, as the choice made by the collective influences everyone in it?

C

Collective Rationality and Social Choice

Of particular interest to the idea of collective rationality is the study of social choice.

In a seminal work, Arrow (1951) connects collective rationality to social choice through the idea of the existence of a social welfare function. Formally, consider a large set of conceivable alternatives, X, that a society faces. A preference ordering R (weakly preferred) on X is a binary relation on X that is both complete and transitive. Its asymmetric and symmetric parts are denoted by P (strictly preferred) and I(indifferent) respectively. There are n number of individuals in the society. Each individual i has a preference ordering R_i on the set X. A *social welfare function* (SWF), F, maps a profile of individual preference orderings $(R_1,\ldots,R_n)$ to a preference ordering on X. The preference ordering $F(R_1,\ldots,R_n)$ is then interpreted as the *society's preference* on X for the society consisting of individuals with preference orderings $(R_1,\ldots,R_n)$. If such an SWF exists, then the social choice to be made from any set of feasible alternatives can be determined by comparing any pair of feasible alternatives according to the society's preference. The social choice thus made is guided by a preference ordering to reach the best among feasible alternatives – *collective rationality* is achieved. In other words, such an SWF, if it exists, is a preference aggregation procedure aggregating individual preference orderings into a society's preference ordering according to which a rational choice can be made.

In isolation, collective rationality is trivial to reach because an SWF always exists. For example, take any preference ordering R on X; the constant function that maps every possible profile of

individual preference orderings to R is an SWF. Obviously, this SWF is not meaningful since no information about individual preferences is reflected by society's preferences. For an SWF to reasonably aggregate individual preferences, some minimal set of conditions should be imposed. Arrow (1951) considers four conditions: U - (universal domain: an SWF's domain contains all possible individual preference orderings), P (Pareto principle: if all individuals strictly prefer one alternative to another, then the society strictly prefers the first alternative to the second), I (independence of irrelevant alternatives: the way the society ranks a pair of alternatives should depend only on the way individuals rank the same pair, not on how they rank any other alternatives), and D - (non-dictatorship: no single individual always gets to determine the society's preference). He shows the famous *Arrow's Impossibility Theorem*: It is impossible to have a social welfare function satisfying U, P, I and D simultaneously. In other words, collective rationality is impossible to achieve universally if society is to take into account all individuals in a minimally reasonable way.

Arrow's Impossibility Theorem jump-started the modern day study of social choice. In the huge literature of social choice theory, two strands directly relate to collective rationality formulated in the context of Arrow's Impossibility Theorem. One strand focuses on identifying domain restrictions so that social welfare functions satisfying Arrow's three other conditions exist. For example, the SWF which derives society's preference from majority voting on each pair of alternatives (majority rule) with universal domain will lead to many cycles in society's preference, violating the transitivity requirement of a preference ordering. However, if individual preferences are restricted to those that are single-peaked when alternatives can be represented in one dimension, then majority rule will not generate cycles and satisfies all other requirements of Arrow's Theorem. In general, this strand of literature proves that collective rationality can be meaningfully restored for some restricted domains (Gaertner 2002). However, domain restrictions are severe, and outside of them the problem of society's preference cycles is global (McKelvey 1979).

The second strand of literature directly examines the formulation of collective rationality in the definition of Arrow's social welfare function. Arrow's SWF requires society's preferences to be orderings, that is, binary relations that are complete and transitive. Suppose that we weaken collective rationality to requiring only that society's preferences be, say, acyclic as opposed to fully transitive. Can impossibility then be avoided? More generally, is the strong collective rationality formulated by requiring society's preferences to be orderings to blame for the impossibility? This line of research concludes that, even with a weakened notion of collective rationality, the impossibility remains (Sen 1995). Therefore, social choice cannot be expected to be collectively rational, even weakly.

Collective Rationality and Strategic Behaviour

The aforementioned work implicitly assumes that truthful individual preferences are aggregated. If, instead, strategic behaviour is allowed, then even if we require a social choice function to be only non-dictatorial (a social choice function maps a profile of individual preferences into an alternative – a choice of the society), every such social choice function can be manipulated. This is the Gibbard–Satterthwaite impossibility theorem (Gibbard 1973; Satterthwaite 1975). That is, even if the collective makes up its mind about what is good for the society in a given circumstance, as long as individuals are free to report their preferences and the collective does not always choose the top alternative of a given agent's reported preference, then the collective's goal cannot be achieved.

Collective Rationality and Aggregate Demand

The disconnection between collective rationality and individual rationality exists in other areas of economics. In consumer demand theory, the Debreu–Mantel–Sonnenschein theorem (Debreu

1974; Mantel 1974; Sonnenschein 1973) states that generally aggregate demand functions do not exhibit any regularity (such as being downward sloping regarding price) even when all individual demand functions are derived from rational decisions in the sense of preference maximization under budget constraints. More specifically, for any given shape of the aggregate demand function (not necessarily downward sloping), there exists a preference profile, one preference for each consumer, such that the aggregate demand function is generated by the individual demand functions derived from that preference profile. On the other hand, empirical evidence suggests that aggregate demand functions often exhibit some regularity even when individual demand functions do not exhibit regular properties from preference maximization under budget constraints (Kirman 2004).

Possible Reconciliation of Individual Rationality and Collective Rationality

The findings in social choice theory and demand theory suggest a fundamental separation between collective and individual rationality. On the one hand, if individuals in a collective are rational, the collective choice is responsive to individuals, and the collective power does not lie in some proper subset of the collective (democratic), then the collective choice cannot be 'collectively rationalized'. On the other hand, in some situations, collective choices can be 'rationalized' even when individuals in the collective do not act as rational individuals. This separation between collective and individual rationality is not unlike Buchanan's critique of Arrow's formulation of collective rationality:

> We may adopt the philosophical bases of individualism in which the individual is the only entity possessing ends or values. In this case no question of social or collective rationality may be raised. A social value scale as such simply does not exist. Alternatively, we may adopt some variant of the organic philosophical assumptions in which the collectivity is an independent entity possessing its own value ordering. It is legitimate to test the rationality or irrationality of this entity only against this value ordering. (Buchanan 1954, p. 116)

The philosophical bases of individualism have many followers in economics. Binmore (1994, p. 142) wrote: 'Game theorists of the strict school believe that their prescriptions for rational play in games can be deduced, in principle, from one-person rationality considerations without the need to invent collective rationality criteria provided that sufficient information is assumed to be common knowledge.' Under the standard assumptions of game theory accounting for individual interests, these game theorists will prescribe that players defect in the Prisoner's Dilemma game. Such play leads to a Pareto-inferior outcome and thus is in conflict with the collective interest. This is not problematic if game theory is a normative theory which prescribes what people should do rationally. However, as a predictive theory it fails to match what people actually play in the Prisoner's Dilemma game. Experimental evidence shows rampant cooperation among players of the Prisoner's Dilemma game (Rapoport and Chammah 1965; Ledyard 1995).

If we make the organic philosophical assumption that a collective is an independent entity, then do we arbitrarily assume a criterion of collective rationality? A more reasonable way of thinking about a collective being as organic is, perhaps, to consider that, in a collective, individuals become social creatures, not mere individuals, and as such their choices have social consequences that they take into account. This can be modelled as individuals' preferences over a given set of alternatives changing depending on whether they are individuals or members of a collective. How preferences are specifically influenced may reflect culture, social convention or custom, so that they are context-dependent. But whatever the cause, this may create sufficient restrictions on the preference domain that collective rationality results as a consequence of some aggregation procedure that is democratic.

See Also

- ▶ Arrow's Theorem
- ▶ Rational Choice and Sociology
- ▶ Rationality, History of the Concept
- ▶ Social Choice

Bibliography

Arrow, K. 1951. *Social choice and individual values,* 2nd ed. New York: Wiley, 1963.

Binmore, K. 1994. *Playing fair: Game theory and the social vontract*, vol. I. Cambridge, MA: MIT Press.

Buchanan, J. 1954. Social choice, democracy, and free markets. *Journal of Political Economy* 62: 114–123.

Debreu, G. 1974. Excess demand functions. *Journal of Mathematical Economics* 1: 15–23.

Gaertner, W. 2002. Domain restrictions. In *Handbook of social choice and welfare*, ed. K. Arrow, A. Sen, and K. Suzumura. Amsterdam: North-Holland.

Gibbard, A. 1973. Manipulation of voting schemes: A general result. *Econometrica* 41: 587–601.

Kirman, A. 2004. The structure of economic interaction: Individual and collective rationality. In *Cognitive economics: An interdisciplinary approach*, ed. P. Bourgine and J. Nadal. Berlin: Springer.

Ledyard, J. 1995. Public goods: A survey of experimental research. In *Handbook of experimental economics*, ed. J. Kagel and A. Roth. Princeton: Princeton University Press.

Mantel, R. 1974. On the characterization of aggregate excess demand. *Journal of Economic Theory* 7: 438–453.

McKelvey, R. 1979. General conditions for global intransitivities in formal voting models. *Econometrica* 47: 1085–1112.

Rapoport, A., and A. Chammah. 1965. *Prisoner's dilemma: A study in conflict and cooperation*. Ann Arbor: University of Michigan Press.

Satterthwaite, M. 1975. Strategy-proofness and Arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of Economic Theory* 10: 187–217.

Sen, A. 1995. Rationality and social choice. *American Economic Review* 85: 1–24.

Sonnenschein, H. 1973. Do Walras' identity and continuity characterize the class of community excess demand functions? *Journal of Economic Theory* 6: 345–354.

Collet, Clara Elizabeth (1860–1948)

R. D. Collison Black

Keywords

Collet, Clara E.; Women's employment

After a period as a schoolteacher in Leicester, Clara Collet became one of Charles Booth's assistants on his Survey of London Life and Labour in 1886. In 1893 she entered the civil service as labour correspondent and later senior investigator for women's industries in the newly established Labour Department of the Board of Trade. The earnings and employment of women became and remained Clara Collet's main concern; her contemporaries recognized her as the principal authority on the subject in Britain. Articles on female labour and earnings were among her contributions to the first edition of Palgrave's *Dictionary of Political Economy* in 1894, and the thorough and lucid reports which she produced on women's industrial employment figured in Parliamentary Papers, contributing to the passing of the original Trade Boards Act of 1906. After her retirement in 1920 from what had by then become the Ministry of Labour, Collet herself served on a number of trade boards, and wrote the section on Domestic Service for the *New Survey of London Life and Labour* directed by her former chief, Sir H. Llewellyn Smith.

The first woman Fellow of University College, London, where she took her MA degree in 1885, Clara Collet was one of the founders in 1890, along with Henry Higgs and H.R. Beeton, of the Economic Club which met there monthly, and acted as its secretary from 1905 to 1922.

She was also a founder member of the British Economic Association, which later became the Royal Economic Society; she served on its Council from 1920 to 1941, and on that of the Royal Statistical Society from 1919 to 1935.

Selected Works

1893–4. Royal Commission on Labour. Employment of women. Miss C.E. Collet. Report. *Parliamentary Papers* 1893–4 [C.6894, XXIII] xxxvii, pt. I.

1894a. Statistics of employment of women and girls. Miss Collet. Report. *Parliamentary Papers* 1894 [C.7564], 1xxxi, Pt. II.

1894b. (With Dora M. Barton.) Female labour. In Palgrave's *Dictionary of Political Economy*, London: Macmillan & Co., Vol. II.

1894c. Females and children, earnings of. In Palgrave's *Dictionary of Political Economy*, London: Macmillan & Co., Vol. II.
1898. The collection and utilization of official statistics bearing on the extent and effects of the industrial employment of women. *Journal of the Royal Statistical Society* 61: 219–260.
1899. Money wages of in-door domestic servants. Miss Collet. Report. *Parliamentary Papers* 1899 [C.9346] xcii.
1931. Domestic service. Chapter VIII in *New Survey of London Life and Labour,* vol. II, ed. H. Llewellyn Smith. London: P. S. King.
1933. Appendix to *The Private Letter Books of Joseph Collet*, *sometime Governor of Fort St. George Madras*, ed. H. H. Dodwell. London: Longmans.

Collusion

Kevin Roberts

Although collusive practices are not restricted to the economic relationships of a well-defined sub-group in society, it is common to use the term collusion in the context of cooperative activity between different firms. With regard to the study of collusion, research has centred on the conditions most conducive to collusion and, in both theoretical and empirical work on the operation of collusive arrangements (see Scherer 1980, Chaps. 6 and 7).

If economic agents are self-interested maximizers then, given that cooperation between a group of firms will almost certainly make possible higher profits for each member of the group than is possible without cooperation, there is a presumption that collusive arrangements will be widespread. Thus it is important to understand why collusion should fail to be universal. In fact, a more general issue is also raised by this. Taking the argument one stage further, direct cooperation between a group of firms and its consumers is likely to make possible benefits to everybody as compared with a situation where firms and consumers are separated by an anonymous market. This cannot be the case for the economy as a whole if it is Pareto efficient (e.g. the standard perfectly competitive economy without externalities) but it will still be the case that cooperative action by a group of agents within the economy will allow that group to gain at the expense of the rest of the economy. The study of the problems faced by colluding firms should hope to throw light on these more general issues.

1. Consider a well defined group of firms producing identical or similar products, the implication being that demand for the products is interrelated. As a first step, assume that cost and demand functions faced by each firm are common knowledge (every other firm in the group knows these functions, every firm knows this, and so on). If the firms meet together to collude then some joint action will emerge. Cooperative game theory concerns itself with this solution but, for the present exercise, it is sufficient to note two of the main determinants of the eventual solution. First, there are the outcomes made possible by cooperative action. Assume profits can be redistributed within the group–side-payments can be made. Then if cooperative action fails to maximize joint profits all firms can be made better off. Given the common knowledge assumption so that there is no argument for inefficiency based upon the mis-perceptions of firms, joint profit maximization, with the group of firms acting like a multi-product monopolist, should emerge.

   Side-payments may not always be possible. For instance, collusive behaviour is outlawed in many countries and, though it may be difficult to detect whether actions by firms are part of some collusive arrangement, the transfer of money between firms is much more likely to be capable of detection. But without side-payments, actions which influence the size of joint profits also influence the distribution of those profits; the consequence of this being that distributional considerations will influence the actions chosen by firms. For instance, the

cooperative Nash bargaining solution (Nash 1950) leads to actions which maximize the *product* of individual gains above some status quo position – compared with joint-profit maximization, there is movement towards the equalization of gains above the status quo. This may be viewed as a compromise between ignoring distributional considerations and the other extreme where only distributional considerations count.

The second major determinant of the solution reached will be what each firm can expect to achieve if it refuses to accept a particular proposed collusive action for the group. The status quo of the Nash bargaining solution may be interpreted in this way. This is the most obvious component of 'bargaining power' for an agent. The requirement that firms must prefer the collusive action to what can be achieved by reneging, places restrictions on the collusive solution. Under the assumption that firms are interrelated only through the demand structure and that this interrelationship implies that the goods produced by the collusive firms are substitutes, then the worst that can happen to a firm if it refuses to accept a collusive arrangement is that all other firms maximize their production and the reneging firm chooses production to maximize profits in this hostile environment. Given this scenario, there will usually be a large range of collusive actions which offer more to firms than can be achieved by reneging. However, it is not enough for firms to say that they will 'punish' a reneging firm in this way, there must be grounds on the part of the reneging firm for believing that punishments will be carried out – it must be a credible threat in the sense that if the firm reneges, other firms have an incentive to punish the firm. The credibility restriction in this environment is captured by the so-called 'perfect folk theorem' of repeated games (see Rubinstein 1979, for a published version of the theorem) – firms called upon to punish will have the strongest incentive to punish if all other firms punish that firm for not punishing in the first place. This argument leads to the conclusion that any collusive outcome can be maintained as long as all firms are better-off than the best they could achieve under maximum retaliation from other firms. Here it should be noted that as, by definition, a firm will be happy to choose its Nash strategy if all other firms do the same, the Nash equilibrium is a feasible punishment solution. Thus there will always be an effective deterrent if all firms are better-off under collusion than in the Nash equilibrium (Friedman 1971).

2. Thus far, the story has assumed that there are no information problems for the (potentially collusive) group of firms. Exogenous uncertainty faced by all firms is not a particular problem but when there are informational asymmetries between firms, the study of collusion is much richer. In fact, almost all the theoretical literature on collusion takes some informational asymmetry as a starting point and it is useful to survey some of this literature from the viewpoint of the informational asymmetry which is being postulated.

   It is convenient to distinguish between two forms of asymmetry – adverse selection where, for instance, some firms do not know the cost and demand conditions of other firms, and moral hazard where it is the behaviour of other firms which cannot be observed. In the former case, it is preferences that cannot be observed, in the latter case it is actions. Both forms of asymmetry can have important effects upon the structure of collusive arrangements that could be expected to emerge.

3. Taking the adverse selection case first, it is fairly reasonable to assume that a firm will have a better knowledge of the demand and cost conditions that it faces than other firms. Although firms may be attracted by the simplicity of adopting a collusive arrangement based upon solely common information, this solution ignores the efficiency gains that may be achieved from making the collusive outcome sensitive to the privately held information of firms. Clearly, the main problem that arises is that each individual firm must have the incentive to reveal this private information. Consider a simple model with just two firms, 1 and 2. Assume that the revenues they receive

when they produce outputs q_1 and q_2 are given by $R_1(q_1, q_2)$ and $R_2(q_1, q_2)$. Each firm has a constant marginal cost of production β_1 and β_2 and assume that this is private information, firm 1 knowing the true value of β_1, firm 2 knowing the true value of β_2. Assume, for simplicity, that β can take on only two values, β and $\overline{\beta}$, $\underline{\beta} < \overline{\beta}$. Collusion will result in the adoption of output levels for each firm and the levels may be sensitive to the private information β_1 and β_2 – $q_1(\beta_1, \beta_2)$ and $q_2(\beta_1, \beta_2)$. However, this can only be implemented if a firm does not pretend to be a high cost firm $(\overline{\beta})$ when it is low cost $\left(\underline{\beta}\right)$ and vice versa; for firm 1, profit maximization implies

$$R_1\left[q_1\left(\underline{\beta}, \beta_2\right), q_2\left(\underline{\beta}, \beta_2\right)\right] - \underline{\beta} q_1\left(\underline{\beta}, \beta_2\right) \geqq R_1\left[q_1(\overline{\beta}, \beta_2), q_2(\overline{\beta}, \beta_2)\right] - \underline{\beta} q_1(\overline{\beta}, \beta_2)$$

and

$$R_1\left[q_1(\overline{\beta}, \beta_2), q_2(\overline{\beta}, \beta_2)\right] - \overline{\beta} q_1(\overline{\beta}, \beta_2) \geqq R_1\left[q_1\left(\underline{\beta}, \beta_2\right), q_2\left(\underline{\beta}, \beta_2\right)\right] - \overline{\beta} q_1\left(\underline{\beta}, \beta_2\right)$$

This places restrictions on the class of solutions that can be implemented. Combining the inequalities gives

$$\left(\overline{\beta} - \underline{\beta}\right)\left[q_1\left(\underline{\beta}, \beta_2\right) - q_1(\overline{\beta}, \beta_2) \geqq 0\right.$$

which implies that the lower the marginal cost, the more a firm is allowed to produce. Given that the collusive solution will usually entail a restriction in output as compared with what a firm would like to choose (the other firm's output remaining constant), the firm must be provided with a disincentive from pretending to be a low-cost firm when it is high-cost and this will come from variations in q_2 – to provide the right incentives to firm 1, the output of firm 2 should be negatively correlated with the marginal cost of firm 1.

The foregoing demonstrates that the implementability requirement gives some structure to the collusive solution. But the exact outcome chosen will be the result of bargaining between the firms concerned. This bargaining process may involve the transfer of information prior to collusive agreements being reached (Roberts 1985).

Recognition of the adverse selection problem gives theoretical insight into the form of collusive practices that are discussed in the more applied literature. In general discussions a distinction is drawn between implicit and explicit collusion. However, in models without information problems, each firm is aware of the agreement that would be decided upon if it met with other firms and is aware how reneging firms would be dealt with. In this case, there is no need for firms to collude explicitly and the distinction between implicit and explicit collusion is not useful. But when adverse selection exists, it is clear that a rule which makes other firms' behaviour depend upon the private information of some firm will require information transmission between firms. The idea of implicit collusion can be rationalized as a situation where either no information is transmitted or, to take a less extreme case, where information is transmitted through aggregate market-wide indices, e.g. the equilibrium price in the industry.

The existence of adverse selection can also provide a theoretical rationale for mark-up pricing as a collusive outcome (Roberts 1983). When firms are selling in the same market, cost conditions facing firms are more likely to be the source of private information than demand conditions. While it may be impossible for other firms to observe the cost *function* of a firm, it may be possible to observe the level of costs at the output being produced. As this observation can be used as a proxy for the private information of the individual firm, a collusive agreement will involve the output and price levels of firms being dependent upon cost levels – this provides a rationale for why the firms' behaviour may be sensitive to average, rather than the conventional marginal, costs.

4. Over the last 20 years, most of the theoretical literature on collusion has considered

situations where the actions of firms fail to be perfectly observed by other firms – a situation of moral hazard. If the preferences of firms are common knowledge then the problem is not one of deciding upon the collusive solution but, instead, of policing that solution.

The simplest example of imperfect monitoring arises when there is a delay before the action of a firm is observed by others. If time units are set equal to the delay time then a firm will gain from reneging on a collusive agreement if

$$(1-\beta)\Pi^R + \beta\Pi^P > \Pi^C$$

where $\prod^C$ is the profit per unit time under the collusive agreement, $\prod^R$ is the maximum profit that the firm can achieve by reneging given that other firms keep to the collusive agreement $\prod^P$ is the profit the firm can achieve when it is being punished by the other firms and β is the discount factor for the firm. Obviously, $\Pi^R \geqq \Pi^C$ so that the longer the delay before other firms perceive reneging (the smaller the β), the more incentive a firm has to renege. As the degree of punishment that can be inflicted is restricted by the requirement that other firms have to punish, and this incentive itself is diminished when there is a delay in observing behaviour, the smallest possible $\prod^P$ will rise as the delay time increases so reinforcing the incentive to renege. For a detailed analysis of this problem, see Abreu (1986).

The other main form of imperfect monitoring that has been considered deals with the case where individual firm behaviour is never observable but market-wide aggregates can be observed by all firms. This was the situation which was studied by Stigler (1964) in his seminal paper on collusion and by many authors since (a recent analysis is to be found in the work of Porter 1983, and Green and Porter 1984). In the case studied by Stigler, firms observe the demand conditions for their own output and this gives an indication of the prices that other firms are charging. Green and Porter take the simple case of a homogeneous product being produced by similar firms. There is some uncertainty in demand so that a 'low' market price may be a result of this uncertainty or of 'cheating' on the collusive agreement by some firms. Green and Porter consider 'trigger-price' punishment strategies which take the form of the group moving to the Cournot equilibrium for a fixed time T if the market price drops below some level $\tilde{p}$. A feature of this informational set-up is that all firms suffer from this punishment. The trigger-price strategies are set so that cheating does not occur though, because of the uncertainty, a proportion of time is spent in the punishment regime. There is a direct trade-off between the gains from a collusive agreement that severely restricts output and the costs of punishment that will be suffered in the maintenance of this agreement. Notice that as there is no adverse selection, there are no requirements for the firms to meet together to decide upon collusive behaviour – it may be far-fetched but there is nothing to rule out a system of implicit collusion with trigger-price strategies. For this reason, these moral hazard models are often described as non-cooperative models of collusion.

5. The theory and practice of collusion are much discussed in texts on industrial organization. The foregoing has tried to make clear that the structure of information asymmetry in the market is crucial for understanding the operation of collusive agreements. With a particular informational set-up, the structure of the set of collusive agreements which will not entail reneging is now quite well understood. Despite much work, there is rather less understanding of the exact agreement that will be settled upon.

See Also

- ▶ Cartels
- ▶ Cooperative Equilibrium
- ▶ Cooperative Games

Bibliography

Abreu, D. 1986. Extremal equilibrium of oligopolistic supergames. *Journal of Economic Theory* 39(1): 191–225.
Friedman, J. 1971. A non-cooperative equilibrium for supergames. *Review of Economic Studies* 38(1): 1–12.
Green, E., and R. Porter. 1984. Noncooperative collusion under imperfect price information. *Econometrica* 52(1): 87–100.
Nash, J. 1950. The bargaining problem. *Econometrica* 18(2): 155–162.
Porter, R. 1983. Optimal cartel trigger price strategies. *Journal of Economic Theory* 29(2): 313–338.
Roberts, K. 1983. *Self-agreed cartel rules*. IMSSS Discussion Paper No. 427, Stanford.
Roberts, K. 1985. Cartel behaviour and adverse selection. *Journal of Industrial Economics* 33(4): 401–413.
Rubinstein, A. 1979. Equilibrium in supergames with the overtaking criterion. *Journal of Economic Theory* 21(1): 1–9.
Scherer, F. 1980. *Industrial market structure and economic performance*, 2nd ed. Chicago: Rand McNally.
Stigler, G.J. 1964. A theory of oligopoly. *Journal of Political Economy* 72(1): 44–61.

Colonialism

M. Abdel-Fadil

> Everywhere do I perceive a certain conspiracy of rich men seeking their own advantage under the name and pretext of the commonwealth. (Sir Thomas More)
> Modern colonialism, as a historical phenomenon of territorial expansion, is intimately entwined with the rise and expansion of the modern capitalist world system. So colonialism is entwined with the history, economics, politics and ruling ideas of the modern capitalist society. On the other hand, and to avoid any terminological confusions, the term *imperialism* should be reserved to designate the new nexus of financial and technological dependency relations and arrangements marking the new distinct stage of mature capitalism. (Magdoff 1970)
> During the modern colonial period (1870–1945) colonialism has emerged as a general description of the state of subjection – political, economic and intellectual – of a non-European society as a result of the process of colonial organization. (Fieldhouse 1981)

The Age of Colonialism: Historical Background

The age of colonialism began about 1500, following the European discoveries of a sea route around Africa's southern coast (1488) and of America (1492). Colonialism thus expanded by conquest and settlement after a period of extensive exploration. The improvement in navigational instruments helped a great deal to make substantial progress in the discovery of new geographical territories.

Portugal emerged as the leading nation in such process of overseas expansion. 'The search for wealth in the form of gold, ivory, spices and slaves spurred the Portuguese and may have been the strongest motivating force behind the colonization drive of the Portuguese during the 16th century' (*Encyclopaedia Britannica* 1768, Vol. 4, p. 881).

The old colonial period, which lasted nearly three centuries, following the major Portuguese and Spanish conquests, may be viewed largely as a commercial venture. The Spaniards and the Portuguese resorted to their warships, gunnery and seamanship to keep the main trade routes open. The Spanish sovereigns created in 1504 the House of Trade (Casa de Contracion) to regulate commerce between Spain and the New World. Their purpose was to establish state monopoly over overseas trade, and thus pour the maximum amount of bullion into the royal treasury (*Britannica* 1768, Vol. 5, pp. 882–3).

The old colonial system was disrupted in the 18th century as new contradictions developed due to the rapid advance of the Industrial Revolution in England, and by the progressive control England was able to exercise over world shipping. Such new developments led to a policy of opening the American ports to international trade, a policy at variance with the type of colonial relations prevailing between Spain, Portugal and their colonies. These relations were organized exclusively around the exploitation of precious metals (Furtado 1970, p. 20).

The century between the 1820s and the outbreak of the World War I saw the establishment of the modern colonial order. For during that period European countries had achieved complete dominance over world trade, finance and shipping. On the other hand, the political and military authority of the European conquerors was backed by superiority in technology, applied science, organization and information systems (Bagchi 1982).

Between the late 1870s and World War I (1914–18), the colonial powers added to their possessions an average of about 240,000 square miles (620,000 sq.km.) a year, while during the first 75 years of the 19th century the rate of increase in new territories acquired by colonial powers averaged about 83,000 square miles (215,000 sq.km.) a year. By the year 1914, the colonies extended over approximately 85 per cent of the surface of the globe.

Against this historical background, John Hicks establishes a useful distinction between two types of colony: colonies of settlement and trading-post colonies (Hicks 1969, p. 51). A third type of colony was identified as 'the plantation colonies'. In such case, the colony which started as a colony of settlement was gradually transformed into a trading colony (ibid., p. 53).

The Colonization Debate

While there is a strong connection between mercantile expansion and colonization, it would be a mistake to emphasize the crude economic interpretation of colonialism by narrowing down colonialism to the process of control of supplies of raw materials, mineral resources and markets in underdeveloped and precapitalist regions. In fact, such a narrow economistic approach eliminates a vital aspect of colonialism relating to political activity and the drive for dominance over the daily lives of the people of the colonized regions (e.g. French colonialism).

Nonetheless, colonialism must be viewed, dialectically, as a complex phenomenon of capitalist expansion, operating in terms of time and space. To illustrate this point, S.H. Frankel described such a process as a *disintegrating* but also a *formative* process, a unique process in the history of mankind (Frankel 1953). Some other writers justified colonial rule on the ground that 'Colonialism was a necessary instrument of "modernization" which would help other peoples to do what they could not have done, or have done as well, by themselves' (Fieldhouse 1981, p. 43). At the other end of the spectrum, radical theorists, notable among them Walter Rodney, claim that under colonialism 'the only things that developed were dependency and underdevelopment' (Rodney 1972, p. 256).

One of the most articulate arguments put forward in defence of the colonial rule in underdeveloped areas is that of Lord Bauer, who contends that:

> The colonial governments established law and order, safeguarded private property and contractual relations, organized basic transport and health services, and introduced some modern financial and legal institutions. This environment also promoted the establishment or extension of external contracts, which in turn encouraged the inflow of external resources, notably administrative, commercial and technical skills as well as capital.
>
> These contacts also acquainted the population with new wants, crops, commodities and methods of cultivation and opened new sources of supply of a wide range of commodities. These changes engendered a new outlook on material advance and on the means of securing it: for good or evil these contacts promoted the erosion of the traditional values, objectives, attitudes and customs obstructing material advance. (Bauer 1976, p. 149)

This argument only confirms the deep-seated Western biased view, claiming that material progress and advance can only be achieved by eroding the traditional values, customs and production structures of pre-capitalist and primitive societies. Rosa Luxemburg (1913) would see in Lord Bauer's view an eloquent proof of her radical contentions about colonialism and territorial expansion, by the emerging capitalist nations.

But what is at issue is not the possibility or not of achieving material progress or advancement, but the terms on which these transformations in the material and socioeconomic structures were operated. From our viewpoint, what needs to be stressed is the loss of sovereignty which the process of European colonization entailed for

practically all colonized peoples. In Africa, for instance, European colonizers often crushed, suppressed or amalgamated states at will. In most instances, the direct colonial rule was designed to direct and reorder the day-to-day lives of the African peoples (Ajayi 1969).

Seen in a radically different light, thinkers such as Albert Memmi, Jean-Paul Sartre and Franz Fanon placed greater emphasis on the ideological implications and the sociopsychological consequences of the process of colonization. According to Fanon, colonialism tended not only to deprive a society of its freedom and its wealth, but of its very character, leaving its people intellectually and morally disoriented (Fanon, English edition, 1966).

Patterns of Colonial Trade

Historians tend to agree that the conquest of colonies was designed to the economic advantage of the European conquerors. Some historians (e.g. E.J. Hobsbawm) would go as far as to claim that the Industrial Revolution in England would not have been accomplished without the conquest and penetration of 'underdeveloped' markets overseas (Hobsbawm 1968, p. 54).

In fact, the primary aim of all European states was to use commercial regulations to maximize their share of colonial trade in both directions and the profits they made from it. The English Navigation Acts, dating from the 1650s, may be taken as typical in this respect. According to these Acts, all colonial trade must be carried in British-owned and registered ships. All goods imported to the colonies must either be the product of Britain or be transhipped and pay duty there. Any colonial exports so 'enumerated' must be carried direct to a British port in the first instance (Deane 1965, p. 204). The aim of such rules and regulations was to give British shipowners, merchants and manufacturers an assured benefit from colonial commerce and to enable the government to tax colonial trade. This clearly indicates the close association between the process of colonization and the rise of various foreign-trade monopolies held by the charted colonial companies.

Hence, against all claims of the 'Free-Trade' school, the British cotton industry did not rely on its competitive superiority, but relied heavily on the monopolistic practices embodied in colonial trade-regulations, and enforced by the British commercial and naval supremacy (Hobsbawm 1968, p. 58). On the other hand, the terms of trade between the colonized areas and the metropolitan countries had a tendency to deteriorate steadily over time, so that the primary producers in colonized areas tended to obtain proportionally less with their labour than they could have done had they concentrated on producing food or other subsistence crops for their own use or for the home market (Fieldhouse 1981, p. 78). This may be characterized, in modern terminology, as 'unequal exchange', which emphasizes once again the exploitative nature of colonial trade.

The Internal Control of the Colonial Economy

The key to understanding colonialism as a historical phenomenon lies in analysing the mechanism of the internal control of the colonial economy. In this connection, one has to answer two fundamental questions. Why did the Western countries spend so much energy, blood and money in seeking to procure colonial possessions? What are the direct and indirect economic benefits of colonialism?

In the colonial economy, top priority was given to infrastructure investment: railways, harbours, telegraphs, rivers and roads, since it was believed that these constitute the prerequisites of a modern economy, making it possible to link internal areas of production to the world commodity markets. In the agricultural sector, it is still an open question whether plantations, owned and run by foreigners, made any significant contribution to the development of the colonial economy. On the positive side, they served as the main vehicle of introducing new crops, attracting foreign capital, expanding the base of the cash economy and the wage-labour force, and increasing agricultural productivity. On the negative side, the crops of such plantations were subject to severe

fluctuations on the international commodity markets, thus subjecting the colonial economy to severe cyclical fluctuations.

In matters of industrialization, many observers tend to agree that the colonial powers did not positively encourage industrialization in their dependencies, and in many instances their basic policies led to some sort of de-industrialization (Bagchi 1982). In this respect, many writers invoke the record of colonization in India from the days of the East Indian Company. For the balance of historical evidence points to the fact that up to the 18th century the economic conditions of India were relatively advanced, and Indian methods of industrial production were comparable with those prevailing in any other advanced part of the world (Baran 1957, p. 144).

On the other hand, Bill Warren has offered a neo-Marxist view, opposed to the 'Dependency School', regarding the effects of colonialism on the development of productive forces in Third World countries. His main argument runs as follows: 'Direct Colonialism, far from having retarded or distorted indigenous capitalist development that might otherwise have occurred, acted as a powerful engine of progressive social change' (Warren 1980).

Nonetheless, Warren's positive account of the effects of colonialism on the process of capitalist development in Third World countries tends to be rather unitary in spirit. For the pattern of resource allocation in colonial territories had been shaped and administered largely by foreign investors, bankers and merchants. According to Paul Baran, the principal impact of foreign enterprise on the development of the underdeveloped and precapitalist regions 'lies in hardening and strengthening the sway of merchant capitalism, in slowing down and indeed preventing its transformation into industrial capitalism' (Baran 1957, p. 205).

This very nature of the process of capitalist development under colonialism led some authors, such as H. Alavi, to offer the highly controversial concept of the 'colonial mode of production'. This concept was offered as a theoretical construction designed to allow for a variety of relations other than those which characterize the 'capitalist mode of production', as experienced in the advanced capitalist economies of the 'centre' (Alavi 1975). In this respect, Alavi and company established the distinctive features of the 'colonial mode of production' on the basis of empirical investigation of the circuits of capital and forms of labour recruitment of what comes to be called by other authors (i.e. Samir Amin and Gunder Frank) 'colonial capitalism' or 'peripheral capitalism' (Booth 1985, p. 169). Yet the difficulties and confusions surrounding the concept of a 'colonial mode of production', as a distinct mode of production, remain formidable.

Decolonization and Neocolonialism: Two Sides of the Same Coin

The drive towards decolonization in the post-World War II period was a response to the economic crisis of an ageing colonial system. This colonial system was found to involve considerable, and sometimes unacceptable, financial costs to the metropolis. Moreover, colonialism had become increasingly discredited among the people of the colonizing nations themselves, just as the emotional strains of suppressing nationalistic movements in colonized regions had become largely intolerable for public opinion (Fieldhouse 1981, p. 24).

Nonetheless, the process of decolonization proved to be a nominal process in the sense that the formal end of colonial rule did not necessarily result in genuine economic independence for the former colonies. There is now a community of view among left-wing economists and writers that decolonization took place when and because foreign monopoly capital felt confident that the colonial society and economy had been so restructured that their interests could be preserved without direct political control. In other words, colonialism had been merely transmitted into perpetual neocolonialism (Baran 1957).

The term 'neocolonialism', which has gained wide acceptance since the mid-1950s, is meant to designate a state of affairs characterized by a structure of dependency relationships whereby the former colonial territories are kept in their

subordinate place within the imperialist system. This is maintained and sustained by means of chronic and structural balance of payments difficulties, arising from the trade, aid and investment relationships with their former or new metropolitan countries (Warren 1973, p. 35).

In sum colonialism may be seen in a historical perspective as one decisive and dramatic stage in the evolution of international economic relationships. The establishment of colonial rule constituted an arbitrary break in the normal course of history, splitting up regions and creating new artificial entities, transplanting new alien values and institutions into colonized societies. One may finally wonder whether it would not have been better for the people of the colonized regions to remain autonomous until certain indigenous forces could gain momentum and generate new conditions for socioeconomic development and material progress.

See Also

► Imperialism

Bibliography

Ajayi, J.F. 1969. Colonialism: An episode in African history. In *Colonialism in Africa 1870–1960*, vol. 1, ed. L.H. Gann and P. Duignan. London: Cambridge University Press.
Alavi, H. 1975. India and the colonial mode of production. In *The socialist register 1975*, ed. R. Miliband and J. Saville. London: Merlin Press.
Bagchi, A.K. 1982. *The political economy of underdevelopment*. London: Cambridge University Press.
Baran, P. 1957. *The political economy of growth*. New York: Monthly Review Press.
Bauer, P.T. 1976. *Dissent on development*. London: Weidenfeld & Nicolson.
Booth, D. 1985. Marxism and development sociology: Interpreting the impasse. *World Development* 13(7): 761–787.
Deane, P. 1965. *The first industrial revolution*. London: Cambridge University Press.
Encyclopaedia Britannica. 1768. 5th ed., vol. 4, 1977.
Fanon, F. 1966. *The wretched of the earth*. New York: Groove Press.
Fieldhouse, D.K. 1981. *Colonialism, 1870–1945: An introduction*. London: Weidenfeld & Nicolson.
Frankel, S.H. 1953. *The economic impact of colonialism on under-developed societies*. Oxford: Basil Blackwell.
Furtado, C. 1970. *Economic development of Latin America*. London: Cambridge University Press.
Hicks, J. 1969. *A theory of economic history*. Oxford: Clarendon.
Hobsbawm, E. 1968. *Industry and empire*. London: Weidenfeld & Nicolson.
Luxemburg, R. 1913. *The accumulation of capital*. London: Routledge & Kegan Paul. 1951.
Magdoff, H. 1970. Is imperialism really necessary? *Monthly Review*, Part I, 22(5):1–14; Part II, 22(6):1–11.
Rodney, W. 1972. *How Europe underdeveloped Africa*. London: Bogle.
Warren, B. 1973. Imperialism and capitalist industrialization. *New Left Review* (81):3–44.
Warren, B. 1980. *Imperialism: Pioneer of capitalism*. London: New Left Books.

Colonies

Donald Winch

The economic advantages and disadvantages of colonies, the best means of establishing them and ensuring their development, and the principles that should govern trade and other relations with the mother country, have persistently served as fertile topics for policy and theoretical debate in the history of political economy. The treatment given here will be confined to the British debate on colonies from the late eighteenth to the first decades of the twentieth century.

The British empire was composed of colonies and ex-colonies which had differing histories of acquisition and varying political and economic relationships with the mother country. It follows that the problems which they posed were equally diverse, as illustrated by the differences between the economies of the British West Indies before and after slavery was abolished, the question of public land disposal and emigration to Canada, Australia and New Zealand, and the tasks of administering an Indian sub-continent with a largely peasant population living close to minimum subsistence levels. To this list can be added the problems of integrating Scotland and Ireland

into the English economy and polity after the respective Acts of Union in 1707 and 1808, where 'colonial' issues – in a technical rather than emotive sense – were often at stake, even if the term was not used to describe them. Indeed, Ireland and India as subsistence farming economies posed similar problems to British administrators, despite major differences in their cultural backgrounds and political status within the empire. For that matter, even the United States after independence could for some purposes be treated as having a 'colonial' relationship with Britain, largely because it remained a major outlet for British capital and labour.

The sheer magnitude of the problems of empire and their changing nature over more than two centuries would guarantee that they bulked large in the minds of British economists. A few strategic examples will show that there has always been a fairly intimate relationship between economics, economists, and empire. Adam Smith may well have advised the imposition of the Townshend duties on North America in 1763, and he was certainly involved in advising the British government on the consequences of American breakaway when these earlier attempts to exert fiscal control led to successful revolt. Malthus held the first Chair of political economy in Britain at an educational institution at Haileybury established to train the servants of the East Indian Company; and James and John Stuart Mill together devoted nearly 40 years to the service of the Company. The younger Mill was also a consistent supporter of schemes involving the 'systematic colonization' of Australia and New Zealand, as well as taking a major interest, along with most of his classical predecessors and successors, in the problems of the Irish economy. The controversy over imperial preference at the turn of the twentieth century underlined the gulf that existed between historical or institutionalist economists and their more orthodox opponents, led Alfred Marshall, who believed that the increasing challenge to Britain's industrial hegemony did not justify abandonment of deductive methods of economic reasoning or those cosmopolitan free trade principles which had spurred British prosperity earlier. Finally, of course, there is Keynes, whose first employment as a civil servant was within the India Office, and whose first major economic work was a treatise on *Indian Currency and Finance* (1913) – a work which attempted to do for its day what Sir James Steuart had done when he wrote *The Principles of Money Applied to the Present State of the Coin of Bengal* in 1772.

A treatment based on chronology and recurring themes seems the best way of dealing with the diversity of Colonies as a topic of economic interest, though it should be remembered that colonies and empire was never treated solely as *economic* problems, even after the inauguration, largely under Marshall's auspices, of a measure of professional distance in these matters.

The initial and longest period in the history of colonial policy began in England during Elizabethan times and effectively ended with the dismantling of what had become known as the 'old colonial system' in the 1820s. This system was loosely based on the amorphous doctrines which Adam Smith subjected to attack in the *Wealth of Nations* as a central part of his condemnation of the 'mercantile system' (later known as Mercantilism). During this mercantile period colonies generated a large body of literature which reflected the overwhelming concern with national power and economic self-sufficiency as the prime objectives of state intervention. Thus colonies not only served direct strategic purposes as naval or military bases, they were also treated as sources of precious metals, and of strategic and other raw materials necessary to Britain's early manufacturing industries – of particular value when carried in British ships and bought through monopolistic arrangements at prices lower than could be obtained on the world market. Colonies were variously regarded as protected export markets, outlets for surplus population, and sources of tribute, in addition to serving occasionally as prison settlements. Although the development of free trade doctrines, together with associated monetary ideas on specie-flow mechanisms, eventually succeeded in undermining many of the arguments in favour of colonial possessions and regulations, economic nationalism and neo-mercantilistic ideas and policies have always exerted a powerful attraction, especially in countries that were

industrial later-comers or anxious to overcome the problems of underdevelopment by means of import-substitution and/or export promotion.

During the eighteenth century, the established wisdom on the subject of colonies came under question largely because the supposed benefits of trade controls to the mother country were connected with the growing cost to Britain of defending and governing her North American colonies. Among the earliest critics of the colonial system from this point of view was Josiah Tucker, who argued that existing benefits in the form of export markets and imported goods would accrue to Britain under free trade, and without the attendant military and economic burdens of empire. Provoked by David Hume's essays on commerce and money, Tucker also engaged in an important dispute with Hume on the question of whether trade relations between rich and poor countries could be considered as equalizing or not, and if so, by what process – perhaps through rising labour costs and prices in the richer trading partner, or through some other mechanisms of stimulus and emulation in the poorer country. The dispute was of relevance to free-trade relations between England and Scotland after Union, and of potential relevance to Britain's relations with Ireland and other 'colonies', whether acknowledged as such or not. Indeed the Hume–Tucker debate was an early example of recognition of the essential similarities between international and interregional trade in a world in which currencies were linked through the gold standard. It also concerned relative rates of growth and the respective merits of agriculture and manufacturing as the basis for a nation's wealth and prospects for economic development. Would those who had the advantages of an early start acquire world dominion and monopoly; and hence would poorer and later starters in the race be forced to employ protective measures to establish and maintain their infant industries? If noticed, the debate would have foreshadowed later issues raised by Friedrich List and Henry Carey with Germany and the United States in mind, as well as other questions such as 'free trade imperialism', the 'permanence' of the dollar problem after World War II, and the debate on 'dependency' and the development of post-colonial underdevelopment in the 1960s and 1970s.

Smith's extensive treatment of colonies in the *Wealth of Nations* (especially Book IV, chapter 7) became the *locus classicus* of the anti-mercantile position, where much of his discussion was interwoven with an account of the founding of the European colonies which brought matters up to the present, namely to the issues underlying Britain's dispute with its American colonies. Whereas Edmund Burke had advocated the relaxation of trade controls and taxation as a means of preserving the political status quo, Smith maintained that Britain's pretensions to empire would remain those of a shortsighted shopkeeper unless some system could be found whereby the debts and current burdens of empire could be shared by the colonies themselves; and he emphasized the point by closing the *Wealth of Nations* with a warning about the potential long-term effects on British growth prospects of existing arrangements. Hence the elaborate scheme he advanced for an imperial (Anglo-American) free trade zone, with provision for complete fiscal harmonization and legislative union. However, since he regarded this proposal as utopian, its purpose was chiefly to underline the precise conditions under which the burdens of empire could be made acceptable. Much the same result could be achieved through free trade and a treaty of friendship, without provision for imperial government.

Smith's close dissection of the various gains and losses involved in maintaining the monopoly of colonial trade employed a quasi-mercantilist idea of 'vent for surplus', as well as other arguments about the effects on profits of colonial markets which could not be squared with later Ricardian orthodoxy on the doctrine of comparative costs, capital accumulation, Say's Law, and the permanent causes of declining profits. Ricardo also pointed out circumstances in which it was possible for the mother country to so regulate the trade of a colony as to make it less beneficial to the colony, and more advantageous to the mother country than free trade – an early version of the terms-of-trade argument for tariffs (not meant for use) which in the hands of Robert Torrens was to

blossom into a case for an imperial *Zollverein* a few years later.

With the support of most political economists, however, the system of colonial preferences was gradually and unilaterally dismantled, beginning with the efforts of Huskisson and Robinson in the 1820s. The views of special interest groups, especially those connected with shipping and the West Indies, which Smith had expected to prevail, were outflanked by an uncertain combination of intellectual argument, political opportunism, and a general realization that Britain's industrial dominance meant that mutual restrictions merely restricted the dominant partner without conferring equivalent benefit. A similar combination involving humanitarian arguments prevailed on the related matter of West Indian slavery and the slave trade. The keystone of Britain's rather isolated status as a free-trading nation was installed with the abolition of the Corn Laws in 1846, a policy that had considerable long-term significance for British agriculture and Britain's relationship to colonial suppliers of food and raw materials, including the United States as well as colonies of recent settlement.

During the 1830s and 1840s public debate on colonies and colonization was dominated by the activities of Edward Gibbon Wakefield and the colonial reformers, a group of radicals dedicated to the revival of 'the lost art of colonization'. Their programme entailed the creation of self-governing colonies as outlets for Britain's surplus capital and labour, avoiding the evils of simply 'shovelling out paupers', abolishing penal settlements, and creating 'civilized' communities enjoying the benefits of free trade and high rates of growth. Wakefield's diagnosis of the simultaneous existence in Britain of surplus capital and labour, and the consequent need for new fields of employment abroad, was developed in opposition to Ricardian orthodoxy on the wage fund and Say's Law. His ideas on the optimal economic development of colonies also conflicted with Smith's view that countries of European settlement enjoying an abundance of land were likely to make rapid economic progress. The key to high rates of growth lay in achieving the correct balance between capital, labour, and their 'field of employment' by restricting access to land. This could be achieved by setting a price on land sufficient to delay dispersal of the wage-labour force, and by using the proceeds of land sale for the purpose of bringing in new immigrants. This policy meant that public land disposal and immigration had to remain an imperial rather than purely local concern, thereby creating scope for conflict when colonies achieved self-government. The Wakefield policy came to be seen as a symbol of imperial oppression, an attempt to place colonial development within a straitjacket designed with European conditions in mind. It also entailed loss of freedom in disposing of one of the main sources of revenue available to self-governing colonies. The colonial reformers' hopes of establishing an empire in which free trade ruled were another casualty of self-government when it led to tariffs being raised by Canada and Australia against British and other goods.

What now seems remarkable is the rapidity and extent of influence exerted over British colonial policy by Wakefield's untried theories, though modern development economics may yield comparable examples. He was also highly successful in convincing a number of leading political economists, not least John Stuart Mill, that his ideas deserved to form the basis for future policy. Mill gave prominence to Wakefield's ideas in his *Principles of Political Economy* and other writings by consistently championing 'systematic colonization' as a solution to Britain's population difficulties; and by treating its application to new countries as a valid exception to the general principle of *laissez faire*, namely as a case where the self-interest principle acting under competitive conditions would lead to a sub-optimal result as far as the community was concerned. As part of his general modification of Ricardo's assumptions concerning capital scarcity and the distant prospect of the stationary state, Mill also endorsed the conclusions of Wakefield's heterodox diagnosis of Britain's economic condition, while denying that it was in conflict with Say's Law and other received Ricardian doctrines. By acknowledging the importance to Britain of the export of capital and labour to colonies, Mill not only removed an obstacle to support for colonization, and hence to

the extension of empire, he opened up a major exception to the comparative cost doctrine as an interpretation of Britain's trading pattern: the trade with colonies now became akin to interregional trade. It should also be noted that Mill, confirming the tradition that the only new economic arguments for protection have been advanced by those who favour free trade as a general rule, gave a cautious endorsement to the infant industry case for tariffs.

India was never a colony in the same sense that North America, Australia and New Zealand were British colonies. Attempts were often made to prevent or discourage European colonization before 1830, and until 1858 India was governed by the East India Company acting on a renewable Charter granted by Parliament. The Company had been subjected to closer government control in the late eighteenth century, deprived of its commercial monopoly in 1813, and finally ceased trading altogether in 1833 when it lost its exclusive privileges over the China trade. These developments represented another victory for the forces of free trade, and they blunted the force of Smith's criticisms both of monopolies and government by trading companies. As we have seen in the case of Tucker, free trade ideas could be associated with a case for complete 'emancipation' (Bentham's term) of all colonies. (It was an association of free trade with anti-imperialism which was an invitation for revisionist historians to counter with a neat, perhaps over-neat, inversion by drawing attention to 'free-trade imperialism'.) But there were fewer spokesman for such ideas in relation to India, where other notions of European superiority and responsibility held sway, along with more mundane considerations connected with the retention of investment, employment, and trading opportunities.

India had provided Smith with a prime example of a stationary or declining state, something that could either be attributed to the deficiencies of its system of government and taxation, or to those of a backward people whose culture constituted a barrier to economic progress, though usually to a combination of both. In addition, criticism from divergent quarters was made of the flow of tribute leaving India for Britain, much of it financed by exports of textiles that competed with domestic industry: Indian commerce came in conflict with British manufacturing interests, which were placated by the imposition of duties on Indian imports. But with the reorganization of the Company, especially after 1813, came new priorities and opportunities for those with ambitions to bring the light of post-Smithian, and more especially, Ricardian political economy to bear on the problems of Indian administration. Such a task proved highly congenial to James Mill, a critic of the Company's monopoly powers who was appointed by the Company in 1819 and rose to the rank of Chief Examiner in charge of political, judicial, and fiscal correspondence with India. It was largely through Mill's efforts, later endorsed by his son, John Stuart Mill, that the Ricardian rent doctrine came to play such a large part in the conduct of Indian affairs. It provided the basis for the *ryotwari* system of land tenure, whereby the state became the sole landlord and met its revenue needs by levying *ryots* or peasant-farmers according to Ricardian principles of pure rent. By confining the state's exactions to rent it was thought that the peasant farmer would enjoy normal profits and wages, and the state would eliminate an intermediary or landowning class of *zemindars* or rent-receivers. The system embodied action according to a clear analytical proposition, antagonism to rent as a form of private income, and a view of the prospects for Indian economic development that treated the peasant proprietor as a capitalistic entrepreneur, freed from the arbitrary exactions of landowners and responding to market incentives – which is not to say that the application of these Western economic ideas to Indian conditions was any more successful than the *zemindari* alternative.

In Ireland, where similar economic conditions of a growing population dependent on a backward agriculture obtained, it proved more difficult to bypass the Irish landowner in order to grant the kind of security of tenure to the peasant proprietor that either existed or was the aim of administrators in India. Indeed, the initial view of economists during the early classical period was unsympathetic to the preservation of peasant

proprietorship in Ireland. The favoured solution was consolidation of tenant-holdings as a preliminary to the creation of a capitalistic form of farming employing agricultural wage-labour along English lines, together with emigration or absorption of the displaced population into alternative employment. Before, but especially after the Great Famine of 1846, emigration, largely unplanned, was the only part of this programme that operated. Under the leadership of John Stuart Mill, J.E. Cairnes, W.T. Thornton and Henry Fawcett, the earlier diagnoses and remedies were entirely recast; a more positive evaluation of the possibilities of transforming cottier tenants into peasant proprietors enjoying security of tenure was registered and advocated as the basis for a solution to Irish problems. It was a position that ran directly counter to English property ideas and involved recognition of the role of custom as opposed to contract in designing policies and institutions for societies that did not conform to the English model.

Mill's deployment of a more relativist approach in policy matters was later to be seen as a welcome, though incomplete concession to a succeeding generation of more full-blooded historical and institutionalist critics of deductive economic theory, with its built-in bias in favour of rational economic man – a creature originally invoked by Mill to underline the contrast with societies where custom prevailed. Such critics were more numerous and vocal after 1870, and the resulting split within the economists' ranks coincided with a campaign for 'fair trade' in the 1880s which blossomed into Joseph Chamberlain's scheme of tariff reform along imperial preference lines – an unwitting return to Torrens's imperial *Zollverein*. The historical economists, led by William Cunningham and W.J. Ashley, had already followed Schmoller and other German exemplars in according a more positive valuation to Mercantilism, and this presaged their endorsement of tariff reform as an imperial remedy for Britain's declining competitiveness. At the price of forsaking free trade, Britain could offer preferential treatment to imperial food and raw materials in return for similar preferences in the markets of her ex-colonies. There had already been a revival of interest in imperial federation, which could be portrayed, as it was by John Shield Nicholson, as a return to Adam Smith's project of empire. While much of this belongs to the larger subject of imperialism, a term which has always carried more nationalistic and ideological oxygen, the episode is chiefly of interest here because it was the occasion for a major challenge to economic orthodoxy. Chamberlain's use of arguments supplied by such economists as W.A.S. Hewins brought Marshall into the professional and political fray with his *Memorandum on Fiscal Policy of International Trade* (1908), a work which is still perhaps the best brief restatement of the free trade position based on a combination of neoclassical trade theory and an empirical analysis of contemporary conditions in the colonies as well as in Britain.

Marshall's pupil, John Maynard Keynes, was not as impressed by his master's memorandum when he looked back on it from the vantage point of 1930. At this time Keynes had decided that a revenue tariff was an acceptable policy for Britain to follow, though not for reasons connected with imperial solidarity. Thus when Neville Chamberlain achieved his father's goal with the passage of the Import Duties Act in 1932, followed by the Ottawa Agreements which began the period of imperial preference, Keynes withdrew his support for tariffs. Nevertheless, any account of the revolution associated with Keynes's name is likely to be incomplete without some reference to the ending of free trade in Britain, coupled as it was with the inauguration of an era in which external monetary constraints on British domestic policy were weakened. Of more long-term significance to imperial policy in the interwar period, however, was the revival of interest in state-assisted settlement in the white dominions and the new emphasis on colonial development, with Africa as well as India now assuming a larger role in official, if not professional, economic thinking. At this point colonies and colonial policy become

something else, the beginnings of modern development economics.

See Also

- ▶ Merivale, Herman (1806–1874)
- ▶ Mill, James (1773–1836)
- ▶ Trade Unions
- ▶ Wakefield, Edward Gibbon (1796–1862)

Bibliography

Ambirajan, S. 1978. *Classical political economy and British policy in India*. Cambridge: Cambridge University Press.
Barber, W.J. 1975. *British economic thought and India, 1600–1858*. Oxford: Clarendon.
Black, R.D.C. 1960. *Economic thought and the Irish question, 1817–1870*. Cambridge: Cambridge University Press.
Black, R.D.C. 1968. Economic policy in Ireland and India in the time of J.S. Mill. *Economic History Review* 21: 321–336.
Drummond, I.M. 1974. *Imperial economic policy, 1917–1938*. London: Allen & Unwin.
Knorr, K. 1944. *British colonial theories, 1570–1850*. Toronto: Toronto University Press.
Semmel, B. 1970. *The rise of free trade imperialism*. Cambridge: Cambridge University Press.
Stokes, E. 1959. *The English Utilitarians and India*. Oxford: Clarendon.
Winch, D. 1965. *Classical political economy and colonies*. London: Bell & Sons.
Wood, J.C. 1983. *British economists and the empire*. London: Croom Helm.

Colquhoun, Patrick (1745–1820)

D. P. O'Brien

Keywords
Colquhoun, Patrick; Education; Unproductive labour

JEL Classifications
B31

Colquhoun was born in Dundee. A successful early career in business led to the position of Lord Provost of Glasgow in 1782 and 1783. In 1789 Colquhoun moved to London and became active as a magistrate. He worked on the provision of poor relief and put forward plans for the reform of London's police. He died in London in 1820.

Colquhoun's interest in poor relief led to his *New and Appropriate System of Education for the Labouring People* (1806), a pamphlet based on his own experience of running a school in Westminster. Like Thomas Chalmers later, he argued for the necessity of education to raise the standards and aspirations of the poor, though primarily in order to curb vice rather than population. This, he believed, was the most cost-effective way of tackling poverty. His *Wealth, Power and Resources of the British Empire* (1814), his last important work, is the one for which he is best known. This contained detailed figures on incomes and occupations and the relative importance of agriculture and manufacturing in Great Britain and Ireland. He also included a history of the public revenue, and descriptive material on the colonies.

The work was not very securely based; McCulloch, who had first-hand experience of trying to construct large-scale statistical data for his *Commercial Dictionary*, was severely critical of it in the *Edinburgh Review* and in *Brande s Dictionary.* But it was followed by later writers and Colquhoun's estimate that unproductive labour, one fifth of the total, received one third of output was widely quoted.

Selected Works

1806. *A new and appropriate system of education for the labouring people.* London: J. Hatchard.
1814. *A treatise on the wealth, power and resources of the British Empire. In Every quarter of the World, including the East Indies; the rise and progress of the funding system explained; with observations on the national resources for the beneficial employment of a redundant population.* London: J. Mawman.

Bibliography

Blaug, M. 1958. *Ricardian economics*. New Haven: Yale University Press.
Deane, P. 1956. Contemporary estimates of national income in the first half of the nineteenth century. *The Economic History Review* 8: 339–354.
Espinasse, F. 1887. Colquhoun, Patrick. In *Dictionary of national biography;* reprinted. Oxford: Oxford University Press, 1973.
McCulloch, J.R. 1835. State and defects of British statistics. *Edinburgh Review* 61: 154–181.
McCulloch, J.R. 1837. *A statistical account of the British Empire; Exhibiting its extent, physical capacities, population, industry, and civil and religious institutions*. London: C. Knight.

Colson, Léon Clément (1853–1939)

R. F. Hébert

French engineer, economist and statistician, Colson was born at Versailles on 13 November 1853, and died at Paris on 24 March 1939. Trained in mathematics as an engineer at the Ecole Polytechnique and the Ecole des Ponts et Chaussées, Colson extended his interests to statistics and economics, eventually teaching the latter at both his alma maters and at the Ecole des Hautes Etudes Commerciales and the Ecole des Sciences Politiques. Despite a lifelong career in the French Ministry of Public Works, he found time to produce several notable works, including his monumental *Cours d'économie politique*.

Colson received high marks for the technical competence of his theoretical exposition. According to Antonelli, his *Cours* rendered the doctrines of the French Liberal School 'scientific'. Divisia hailed it as the best work on pure theory since Walras. Colson particularly demonstrated his competence in the field of production by erecting a theory of full employment on the idea of capital–labour substitution and the equalization of factor returns at the margin. He followed the Austrian theory of value, but at the same time showed more sympathy to Walras than his contemporaries did. His work also had a certain affinity with Marshall's in that it integrated mathematics and geometry into the exposition of economic theory. As with Cheysson, Colson's economic views were affected by his statistical studies as much as the latter were affected by the former. Despite exceptional difficulties, he was one of the first to attempt an estimate of French national income in the early years after World War I.

Colson was elected to the Institut International de Statistique in 1906, to the Academie des Sciences Morales et Politiques in 1910 (replacing Cheysson), and to the Conseil Supérieure de Statistique in 1912. He belonged to the French Legion of Honour and the Société d'Economie Politique, of which he was president from 1929 to 1933. He became an honorary fellow of the Royal Statistical Society (London) and was elected an original fellow of the Econometric Society, founded in 1931. His most imposing legacy, however, was the generation of 20th-century French economists and engineers he trained at the *grandes écoles* during the interwar period.

Selected Works

1901–7. *Cours d'économie politique*, 3 vols. Paris: Gauthiers-Villars.
1912. *Organisme économique et désordre social*. Paris: E. Flammarion.
1914. *Railway rates and traffic* (trans: Christie, L. R., Leedham, G. and Travis, C.). London: George Bell & Sons.

References

Divisia, F. 1950. *Exposés économiques. L'Apport des ingénieurs français aux sciences économiques*. Paris: Dunod.
Flux, A.W. 1939. Clément-Léon Colson. *Royal Statistical Society Journal* 102(4): 624.
Marshall, A. 1933. Alfred Marshall, the mathematician, as seen by himself. *Econometrica* 1: 221–222.

Colwell, Stephen (1800–1871)

Henry W. Spiegel

Stephen Colwell, American protectionist, was born in Virginia (now West Virginia). After practising law, he eventually became a successful industrialist and entrepreneur in Philadelphia, where he was a leading citizen and philanthropist. He was a friend of Henry Carey's and shared many of Carey's views, especially his ardent protectionism. Colwell's appeal for high tariffs on iron manufactures and other goods resounded in many publications. Some of these were addressed to the Presbyterian clergy and he drew on religion to fortify his economic views. Colwell buttressed his appeal by making it part of a wider view of the world that may be characterized as elitist and supportive of high wages but also of inequalities of wealth and status. These were to be offset by a stewardship of wealth, that called for private charity rather than public relief for the poor, which Colwell opposed. High wages and private charity thus became complements of high tariffs.

Colwell's arguments ran counter to the teachings of the classical economists, whom he criticized on grounds that were similar to the criticisms made by the exponents of German historical economics and economic romanticism. Colwell was impressed by the protectionist views of Frederick List and found a translator for the original German of the latter's *National System of Political Economy*, to which he himself wrote an introduction.

Colwell is also remembered as the author of *The Ways and Means of Payment*: *A Full Analysis of the Credit System with its Various Modes of Adjustment* (1859), a massive volume that treats of the financial controversies of the time. In this work Colwell supported a private national bank, inconvertible paper money, the real-bills doctrine, the demonetization of gold, and a national clearing system, all amidst an economy of high prices. He denied the validity of the quantity theory of money and advised against a 100% reserve plan. He wrote the book from a point of view that considered money the handmaiden of commerce.

Selected Works

1851. *New themes for the Protestant clergy.* Philadelphia: Lippincott, Grambo & Co.

1856. Preliminary essay to List, F. *National system of political economy.* Trans. by G.A. Mantile, Philadelphia: J.B. Lippincott & Co.

1859. *The ways and means of payment: A full analysis of the credit system with its various modes of adjustment*. Philadelphia: J.B. Lippincott & Co.

Bibliography

Dorfman, J. 1946. *The economic mind in American civilization 1606–1865*, vol. 2. New York: Viking.

Mints, L.W. 1945. *A history of banking theory in Great Britain and the United States*. Chicago: University of Chicago Press.

Combination

John Saville

'Combination' is a term used for a variety of forms of organization. An obsolescent usage relates to business firms which have come together in some kind of merger and today are usually referred to as monopoly, cartel, industrial combination or multinational. In Britain during the 18th century and for most of the 19thcentury combination was understood to mean associations of working men whose purposes were the raising of wages or the alteration of working conditions. The term 'trade union' did not come into common use until after 1830 and only in the second half of the century did it supplant 'combination'.

Combined labour action in the 18th century was widespread although by no means was it all directed by or channelled through formal organizations. The Webbs (1894, ch. 1) were in error in insisting that only those associations that were formally constituted and in continuous existence should properly be counted among the early trade unions. In the 18th century, especially, it was the collective presence of workers that was the crucial determinant of industrial response. Formal organization was not a necessary condition of industrial militancy, and 'collective bargaining by riot' is a well-documented phenomenon (Hobsbawm 1964, ch. 2). In 1718 and 1724 West of England clothiers complained to Parliament that weavers had 'threatened to pull down their houses and burn their work unless they would agree to their terms'. A study of Lancashire textile workers (Turner 1962) emphasized that the essence of 18th-century unionism was the persistence of collective pressures which in given circumstances encouraged collective action. Associations among the Lancashire textile workers developed informally out of the occupational life-style within the community. A settled group would encourage habits of association and common action, and skill was always a consolidating factor, but there was also much activity among the unskilled and migrant groups.

An incomplete listing of recorded industrial disputes gives a total of 383 for the whole of Britain between 1717 and 1800 (Dobson 1980, ch. 1). Most of these were in England, and of the English figure of 333 disputes just over a third (120) occurred in London, the main centre of the artisan trades in the 18th century. The occupational breakdown for the whole country shows 64 incidents from among the textile workers, mostly in the wool industry; only seamen and ships-carpenters (each with 37) and tailors (with 22) exceed a total of twenty. Those between ten and twenty include coalminers, workers in the shipbuilding industries, textile workers other than those in wool or silk, shoemakers, and most trades in building. The range of additional trades for which some disputes were listed is considerable, and hardly any occupation except the service industries was free from industrial conflict (Rule 1981, ch. 6).

The London journeymen tailors were the most effective combination in Britain during the 18th century. Their organization seems to have been established around 1700 as the result of the coming together of five 'box' clubs: box clubs being a version of the friendly society. In order to function effectively the clubs were associations in continuous existence, and their rules and regulations later provided the basis for the discipline of a trade union. The meeting places of the box clubs were public houses recognized by the trade and known as the 'house of call'. The box club and the house of call were the type of organization common to all the trades that succeeded in establishing more or less continuous associations. The tailors' combination first came into prominence as the result of a petition presented to the House of Commons in 1721, which asserted that 15,000 London journeymen had entered into a combination and engaged upon a strike. The report of a House of Commons committee was followed by an Act (7 Geo 1, c. 13) which fixed wages by the day for summer and winter, and which also prohibited combinations. While the Act was going through Parliament the journeymen briefed counsel, at a reported cost of £700.

Legislation against combinations in specified trades was common during the next hundred years; examples affect the wool trade, (12 Geo, I, c. 34), hatters (22 Geo II, c. 27), silk weavers (17 Geo III, c. 55), and the paper trade (36 Geo III, c. 111). At the end of the century, in 1799, there was passed a general Act (to be repealed and replaced the next year by 40 Geo III, c. 60) which made illegal all combinations whose purpose was obtaining an advance in wages or the lessening in the number of hours worked; and for the next quarter of a century the Combination Laws were in force, although in respect of certain trades they were not always rigorously applied (Webb and Webb 1894, ch. 2). In 1824 all previous statutes in respect of combinations were repealed but this Act was in turn repealed and replaced the following year by 6 Geo IV, c. 129, which permitted combinations

on strictly defined terms and listed punishments for the use or the threatened use of intimidation, molestation or violence in the pursuit of the declared objects of the industrial action.

The ending of the complete legal prohibition of combinations was largely due to a growing appreciation of the requirements of a labour market in the period of early industrialization. It was also in part the result of adroit political pressure by Francis Place and, inside Westminster, Joseph Hume. The most weighty supporter of those who accepted the demand that workingmen's combinations should be free of legal restraint was J.R. McCulloch, who spoke in the name of orthodox political economy. McCulloch argued that the Combination Laws were unjust in that employers and their workmen were not put on the same level; they were dangerous because they engendered contempt for the law and encouraged class hatred; and they were futile because no action could permanently drive up the level of wages above the natural rate. McCulloch believed in peaceful combination: 'There is no good reason why workmen should not, like the possessors of every other valuable and desirable article, be allowed to set whatever price they please upon the labour they have to dispose of' – but he was clear above all that no artificial levels of wages could possibly maintain themselves in a competitive market. He saw the usefulness of combinations in a strictly narrow context. He warned against strikes, as normally benefitting those employers outside the strike action, and underlined the danger of industry-wide strikes in reducing the competitiveness of home industry compared with foreign production (O'Brien 1970, pp. 366–70). He was notably adamant against any use of force or intimidation in the day-to-day activities of the combinations and was particularly opposed to attempts to compel workmen to join combinations or participate in strikes or other action; and he commended the clauses in the 1825 Act which imposed prison sentences for those convicted of such intimidation. McCulloch elaborated his ideas in the *Essay on Wages* of 1826, which he published in a revised version in the *Treatise on Wages* of the early 1850s, and he summarized his views in an article on 'Combination' in the 8th edition of the *Encyclopaedia Britannica* of 1854.

The character of trade unionism changed in certain respects during the second quarter of the 19th century as a result of the influence of Owenism and the ideas of the anti-capitalist theorists such as Hodgskin, Gray and Thompson (wrongly designated as the 'Ricardian socialists'). It is important not to exaggerate the nature of the change since unions remained what they had always been: defensive-offensive bodies concerned with the betterment of their members. But in the decade after the legalization of combinations for peaceful agitation there was a notable growth of cooperative organizations and some support for Owen's communitarian ideas. Radical political economy worked in the same direction. The distortion of exchange values, so it was argued, meant that labour exchanged below and commodities above their natural values, and under-consumption and the usual accompaniments of economic crisis were the result. Hence a concern with equitable exchange relations with an emphasis upon the Labour Exchanges of the early 1830s and a longer term preoccupation with money and the banking system (N.W. Thompson 1984). Owenism, with its ideas of cooperation was a central strand in the organization of the Grand National Consolidated Trades' Union in 1834–5 from which, it should be noted, a number of skilled societies held aloof; and Owenism, which was critical of policies such as mutual support funds for unemployment, sickness and death, and remained only a partial influence upon union attitudes. After the failure of the Grand National in the mid-1830s the combinations which survived – almost entirely made up of skilled workers and craftsmen – continued with their traditional sectional approaches to industrial problems. There remained a residue of Owenite ideas – in cooperative production for example which the Christian Socialists of 1848–1854 were able to draw upon (Raven 1920; Saville 1954) – but these were attitudes which became progressively weaker in the decades after the half century.

Combinations of working men for industrial purposes generated a continuous opposition, amounting to hatred, among the employing classes, whose distrust and dislike goes back to the early beginnings of modern industrial society. This hostility, as well as fear, has been a component part of the English liberal tradition and has influenced politicians, administrators and the judiciary. The last group is especially important in that while politicians have slowly modified their views and gradually legislated in more sympathetic ways in respect of trade union rights and status – at least until the last quarter of 20th century – the judiciary have been more wayward in their judgements and have followed quite closely the vagaries of middle class opinion. This opinion has moved between reluctant acquiescence and straight hostility, and the history of the law relating to combinations and trade unions has exhibited batches of legal decisions which have negated the intentions of Parliament. This was true of the late 1860s and during the 1890s, the latter period culminating in the Taff Vale decision. The hostility of educated opinion towards these industrial associations of working men was well illustrated in a letter written by Richard Cobden, the quintessential middle-class liberal of the 19th century, addressed to his brother in 1844: 'Depend upon it, nothing can be got by fraternising with the trade unions. They are founded upon principles of brutal tyranny and monopoly. I would rather live under a Dey of Algiers than a Trade Committee' (Morley 1881, p. 299). These have been enduring sentiments.

See Also

► Industrial Relations

Bibliography

Dobson, C.R. 1980. *Masters and journeymen. A prehistory of industrial relations 1717–1800*. London: Croom Helm.

Hobsbawm, E.J. 1964. *Labouring men*. London: Weidenfeld & Nicolson.

Morley, J. 1881. *The life of Richard Cobden*, Abridged ed. London: Nelson's Shilling Library, 1910.

Musson, A.E. 1972. *British trade unions, 1800–1875*. London: Macmillan.

O'Brien, D.P. 1970. *J.R. McCulloch: A study in classical economics*. London: Allen & Unwin.

Raven, C.E. 1920. *Christian socialism 1848–54*. London: Macmillan.

Rule, J. 1981. *The experience of labour in eighteenth-century industry*. London: Croom Helm.

Saville, J. 1954. The Christian socialists of 1848. In *Democracy and the labour movement*, ed. J. Saville. London: Lawrence & Wishart.

Thompson, E.P. 1963. *The making of the English working class*. London: Victor Gollancz.

Thompson, N.W. 1984. *The people's science. The popular political economy of exploitation and crisis, 1816–34*. Cambridge: Cambridge University Press.

Turner, H.A. 1962. *Trade union growth, structure and policy. A comparative study of the cotton unions*. London: Allen & Unwin.

Webb, S., and B. Webb. 1894. *The history of trade unionism*. London: Longmans & Co.

Combinatorics

A. P. Kirman

Combinatorics, or combinatorial mathematics, is a difficult field to define. It cuts across many branches of mathematics yet a mathematician will clearly sense which problems are of a combinatorial nature. Perhaps the simplest definition is that it is concerned with configurations or arrangements of elements, usually finite in number, into sets. Three basic types of problem are posed. Firstly the existence of certain configurations; secondly, once their existence is proved, the classification or enumeration of the configurations meeting the requirements imposed; and thirdly the construction of algorithms for finding the configurations in question.

Why has combinatorics been the poor relation of the mathematical tools used in economics? The first and most obvious explanation is that the evolution of theoretical or mathematical economics has led us in the opposite direction to that in which combinatorics is useful. Ever since the

'marginal revolution' there has been clear emphasis on the use of differential methods and an implicit acceptance of the perfect divisibility of goods. Indeed if we consider the most elaborate extension of the basic Arrow–Debreu model it is to one in which there is a continuum of agents and a continuum of goods. This is just the sort of context in which the combinatoric approach is of little use.

Yet the economist may well feel that discrete problems are of importance and that indeed the world is best represented as one where goods are not infinitely divisible and where the number of agents is finite. Now given the standard assumptions of convexity, at least in production, the perfectly competitive model may indeed be regarded as a satisfactory ideal or limiting case and the finiteness of a real economy is just an inconvenience. In this case it would seem that combinatorics has little role to play.

However although we may consider taking divisibility as a reasonable idealization in a large economy (even this may be questioned – *see* ▶ Indivisibilities) as soon as convexity is dropped as an assumption for production the problem of finiteness cannot be avoided. If there are increasing returns to scale there will be a minimum profitable plant size and there will be a fundamental indivisibility. At this point combinatoric analysis comes back into its own, and no argument can be made for using infinitely divisible goods as a reasonable approximation. Thus the existence of non-convexities will lead us back into a situation which may, for example, be game-like and in which we will be looking for a solution with a finite number of large plants.

This is to suggest that realism may lead us away from the smooth differential world to which economists are accustomed towards a discrete one in which combinatorics plays an important role.

Nevertheless, till now, finiteness and the combinatoric approach have not occupied a significant place in economics. However, some examples will show that certain branches, although not central, have made extensive use of such an approach.

The development of mathematical programming, in particular of linear programming, has relied particularly on combinatorics. The algorithms developed to solve such problems are essentially combinatoric. Many economic models have been built on the basis of the 'activity analysis' or fixed coefficient production approach.

Game theory has made extensive use of combinatorics and provides an interesting example of how the combinatoric approach can be confronted with one using continuous functions or compact sets.

Combinatoric arguments are used to show that the core of a balanced game is non-empty. A simple argument shows that a market game is balanced and further it can be shown that the only allocations remaining if we replicate a given economy are competitive. This leads us to the conclusion that a competitive or Walrasian allocation exists. Now the latter statement is known to be equivalent to the existence of a fixed point for a continuous mapping. Thus we arrive by combinatorical methods at the same result as that obtained by an apparently very different tool.

This argument is reinforced by the fact that the algorithms developed for finding approximate competitive equilibria are essentially combinatoric in nature. They consist in systematically examining points in the price space, of evaluating the total excess demand of an economy at these points and finding a path which leads to a reduction in the excess demand until it is close to zero. Again, the approximation of fixed points of countinuous functions is obtained by a combinatoric approach.

There are many other examples of ways in which combinatoric analysis has proved useful. Arrow's theorem on the impossibility of a social welfare function is typically proved in the case of a finite number of individuals faced with a finite number of social alternatives. The problem is to find a way of aggregating individual preference orders on these alternatives into a social order in a way which satisfies certain simple axioms. The usual line of proof consists of finding certain configurations of individual preferences which

lead via the axioms to a contradiction with the existence of a social order. The reasoning here too is combinatoric. The introduction of graph theory to describe communication patterns in economics has brought into economic theory a field which has long been regarded as fundamentally combinatoric. The use of 'matching' models to analyse job search and unemployment is yet another example.

Whilst these few rather arbitrary examples show that combinatorial analysis has not been absent from economic theory it is also clear that it has not been central.

However, it seems likely that economics is to evolve towards the sort of models now widely studied in computer science, away from global optimization towards more simple forms of 'rationality' as embodied in simple automata. Furthermore the computation of equilibrium even for underlying continuous models is becoming increasingly important. All this together with a recognition of certain fundamental indivisibilities in economics is likely to move combinatorics to a much more central position on the stage of economic theory.

See Also

- ▶ Cores
- ▶ Fixed Point Theorems
- ▶ Game Theory
- ▶ Graph Theory
- ▶ Ramsey, Frank Plumpton (1903–1930)

References

There elegant and elementary introductions to combinatorics are Berge (1971), Polya, Tarjan and Woods (1983) and Ryser (1963). A more advanced text is Aigner (1979).

Aigner, M. 1979. *Combinatorial theory*. Berlin: Springer.

Berge, C. 1971. *Principles of combinatorics*. New York: Academic.

Polya, G., R.E. Tarjan, and D.R. Woods. 1983. *Notes on introductory combinatorics*. Boston: Birkhauser.

Ryser, H.J. 1963. *Combinatorial mathematics*, Mathematical Association of America. New York: Wiley.

Command Economy

Richard E. Ericson

Abstract

The concept of a 'command economy', a construct in the theory of comparative economic systems, is defined, and its origins, characteristics, and consequences for any society in which it is implemented are explored. The impossibility of the absolute centralization which it requires generates compromises with the market forces it aspires to replace, fostering a symbiotic marketized 'second economy' which systematically undermines its foundations. Hence, although initially appearing to be a true alternative to the market economy, a command economy, most nearly realized in the Soviet Union (1930–87), proved to be ultimately non-viable, collapsing under reforms attempting to make it competitive with market systems.

Keywords

Active vs passive money; Aggregation; Balance; Bounded rationality; Bureaucracy; Central planning; Centralization; Command economy; Command mechanism; Command principle; Communism; Complex social economy; Corruption; Decentralization; Discretion; Gorbachev, M. S.; Gross output; Inca production system; Incentive provision; Industrialization; Inequality of income; Information; *khozraschet*; Labour discipline; Market mechanism; Market versus plan; Microbalance; Moneyness; Mormon economic system; Neurath, O.; *perestroika*; Price control; Principal and agent; Rationing; Resource allocation; Second economy; Sellers' market; Shadow economy; Soft-budget constraint; Soviet economic reform; Soviet Union; Stalin, J.V.; Suboptimization; Unit of measure; Vested interests; War and economics; War communism

JEL Classifications
P3

A command economy is one in which the coordination of economic activity, essential to the viability and functioning of a complex social economy, is undertaken through administrative means – commands, directives, targets and regulations – rather than by a market mechanism. A complex social economy is one involving multiple significant interdependencies among economic agents, including significant division of labour and exchange among production units, rendering the viability of any unit dependent on proper coordination with, and functioning of, many others.

Economic agents in a command economy, and in particular production organizations, operate primarily by virtue of specific directives from higher authority in an administrative/political hierarchy, that is, under the 'command principle'. Thus the life cycle and activity of enterprises and firms, their production of output and employment of resources, adjustment to disturbances, and the coordination between them are primarily governed by decisions taken by superior organs responsible for managing those units' roles in the economic system. One of the most distinctive features of such an economy is the setting of the firm's production targets by higher directive, often in fine detail. The administrative means used include planning, material balances, quotas, rationing, technical coefficients, budgetary controls and limits, price and wage controls, and other techniques aimed at limiting the discretion of subordinate operational units/firms. The command principle strives to fully and effectively replace the operation of market forces in the key industrial and developmental sectors of the economy, and render the remaining (peripheral) markets manipulable and subordinate to political direction. Thus the command principle is likely to clash with the operation of market forces, yet a command economy may nonetheless contain and rely on the market mechanism in some of its sectors and areas, for example, influencing labour allocation, or stimulating small-scale private production of some consumables.

The term 'command economy' comes from the German *Befehlswirtschaft*, and was originally applied to the Nazi economy, which shared many formal similarities with that of the Soviet Union. It has received its fullest development in the analysis of the economic system of the Soviet Union, particularly under Stalin, although it has been applied to wartime administration of the US economy (1942–6; see Higgs 1992), the Mormon economic system in mid-19th century Utah (Grossman 2000), and the Inca production system in the 16th century Andes (La Lone and La Lone 1987). Synonymous or near-synonymous terms include 'centrally planned economy', 'centrally administered economy', 'administrative command economy', 'Soviet-type economy', 'bureaucratic economy' and 'Stalinist economy'.

The command economy's conceptual origins go back to the Viennese economist Otto Neurath, who in the years before and after the First World War developed an extreme version (to the point of moneylessness) based chiefly on prior experience with wartime economies (Raupach 1966). The concept of the command economy has since become a central conceptual framework in the analysis of economic systems, as it captures a logically coherent alternative to 'the market' as a way of organizing socially complex economic activity and interaction. The Soviet Union provided the most complete, and for a while successful, example of a command economy as a working alternative to a market system. Indeed, apart from the relatively short-lived Nazi case, and even briefer ones under emergency conditions in some other countries, especially in wartime, actual instances of command economies are virtually limited to Communist-ruled countries, with the USSR as the prototype and prime exemplar. Thus, what follows is mainly inspired by the Soviet example (Ericson 1991) as it existed, essentially little altered since its appearance in the 1930s, until its collapse in the aftermath of President Gorbachev's *perestroika*, begun in 1987.

Nature of the Command Economy

The seminal analysis of the nature, characteristics, and problems of a command economy is Grossman (1963).

Any complex social economy must, for its very survival, maintain at least a 'tolerable' micro-balance, 'that minimal degree of coordination of the activities of the separate units (firms) which assures a tolerably good correspondence between the supply of individual producer and consumer goods and the effective demand for them' (Grossman 1963, p. 101). In such an economy, appropriate balance can be achieved through decentralized, market-based (monetized, price-mediated) interaction of autonomous units, or by virtue of explicit specific coordinating directives (commands, targets) from some higher authorities. While the former is characteristic of a market economic system, the latter is defining of a 'command economy'. In the latter operational-level units (for example, firms) must merely 'implement' commands; they become 'executants' of plans and directives from above, plans which must insure balance through the coherence and consistency of the instructions they give. Thus the command mechanism requires relative centralization and severe restriction on the autonomy of subordinate operational units. It derives from the overwhelming priority of social goals, and requires the severe limitation, if not total destruction, of autonomous social and economic powers and the enforcement of strict obedience to directives.

A command economy is hence a creature of state authority, whose marks it bears and by whose hand it evolves, exists and survives. Command economies are imposed, whether through external duress or imitation, or indigenously in order to achieve specific purposes such as (a) maximum resource mobilization towards urgent and overriding national objectives, such as rapid industrialization or the prosecution of war; (b) radical transformation of the socio-economic system in a collectivist direction based on ideological tenets and power-political imperatives; and (c), not least, curing the disorganization of a market economy brought about by price control, possibly occasioned by inflationary pressure arising from (a) and/or (b).

The command economy therefore requires a formal, centralized, administrative hierarchy staffed by a bureaucracy, and it also needs to be embedded in (at least) an authoritarian, highly centralized polity if it is not to dissolve or degenerate into something else. And that bureaucracy, if it is to effectively implement the command principle, must exercise full control and discretion, if not necessarily formal ownership, with respect to the creation, use and disposal of all productive property and assets. At the same time, each office or firm and every economic actor within the command structure holds interests which, if only in part, do not coincide with those of superiors or of the overall leadership. This generates important problems of vested interests, principal–agent interaction, incentive provision, and general enforcement of the leadership's will, and calls for a variety of monitoring organizations (party, police, banks, and so on). The term 'command' must not be taken to preclude self-serving behaviour, bureaucratic politics, bargaining between superiors and subordinates, corruption, peculation and (dis)simulation. On the contrary, such behaviour tends to be widespread in a command economy; yet the concept of a 'command economy' remains valid so long as, in the main, authority relations and not a market mechanism govern the allocation of resources.

When not externally imposed, command economies typically arise from a millennialist elite, with unique access to 'the truth', achieving the political power to impose its will, while facing a crisis of apparently overwhelming proportions. The perception of a life-threatening crisis, driving the need for massive mobilization of all social resources and rendering potentially disastrous any hesitation or dissent, any questioning of ways and means, naturally leads, pushed by the 'logic of events', to the usurpation of all power of discretion, all legitimate authority, by the 'knowing' elite, which then becomes responsible for all that is done or not done in the society and the economy. The crisis may be artificial or real ('hostile encirclement'), externally or internally imposed (the need to industrialize, to 'catch up'),

but it requires moving resources rapidly and massively, forcing new activities and interactions in the face of severe scarcities, of shortage of competent personnel, of massive uncertainties, and of strongly held, stark priorities. Indeed, a sense of overwhelming urgency and need for haste drove the elite of the Soviet Union in the 1930s to test and establish, through trial and error over several decades, the institutional structure of a 'command economy', albeit *less than absolute* from both necessity and choice (for example, due to the 'lessons' of 'War Communism') (Grossman 1962; Zaleski 1968).

Consequences of Command

Rational application of the command principle calls for planning, which is basically of two types. Longer-term, developmental planning expresses the leadership's politico-economic strategy (for example, five-year and 'perspective' plans); shorter-term, coordinative planning (annual, quarterly, monthly, ten-day) ideally translates the strategy into resource allocation while aiming to match resource requirements and availabilities for individual inputs, goods, and so on, in a sufficiently disaggregated way for given time periods and locations. The task of elemental coordination, of microbalance, so effortlessly accomplished by any functioning (however poorly) market system, is overwhelmingly large, and grows rapidly with industrialization and economic development, both of which lead to exponential growth of the complexity of the economy, and hence of the planning problem. With centralization and the abandonment of markets comes the need for massive, detailed coordinative planning, for 'making ends meet' in the expanding web of interconnections that must be maintained for economic life to continue. Coordinative planning serves, therefore, as the basis for specific operational directives to producers and users, thereby implementing the command principle to achieve the prime imperative of a social economy – 'balance'.

> It is *this* task that in fact consumes the largest part of the so-called planning in the command economy . . . Coordinative planning as it is conducted in the Soviet Union does little by way of consciously steering the economy's development or finding efficient patterns of resource allocation. Its overwhelming concern is simply to equate both sides of each 'material balance' by whatever procedure seems to be most expeditious (Grossman 1963, p. 108).

A major problem is that detailed planning and the corresponding directives are often late, are insufficiently detailed, may lack the requisite information, hence often cannot be effectively coordinated, and owing to their rigidity are peculiarly vulnerable to uncertainty (Ericson 1983). Information in the command sector, by the logic of the system, tends to flow vertically up and down the administrative hierarchy rather than horizontally between buyer and seller, adding to difficulties of demand–supply coordination by informationally isolating operational units from their suppliers and users. In addition, problems of motivation, accountability (down as well as up), inappropriate decision-making parameters, and divergent interests complicate the procedure. Even at best, this manner of resource allocation can hope to attain only internal consistency (in the sense of effectively matching partially disaggregated requirements and availabilities) but not a higher order of economic efficiency. Economic calculation in pursuit of efficiency enters, if at all, at the project-planning stage, and not short-term resource allocation and use.

These problems are aggravated by the logic of haste that drove imposition of the command economy – 'the pressing contrast between urgent political goals and available resources' (Grossman 1963, p. 108). The necessary attention to the growing problem of balance further militates against any effort to consider developmental objectives or efficiency in making allocative decisions, so that a further bias against allocative efficiency is built into the command economy. Coupled with limited ability to gather, filter, process, and communicate information, and to compute solutions to planning problems, this creates a fundamental and growing inability to acceptably solve the underlying coordination problem, and hence further undermines any consideration of efficiency.

The logic of 'command' has a number of other consequences reflected in the institutions of such

an economy. Planning in a command economy must be largely in *physical terms* due to the crucial importance of balance. The bottom line of the planning process must be available physical units of required inputs, in appropriate assortment, quantity and timing, necessitating physical targets for production and input utilization. Thus tens of thousands of materials and equipment balances must be drawn up and coordinated for each plan period, and then broken down and allocated in directives to specific implementers. And, to be directly usable, these must be in physical or crypto-physical (constant price) units that directly relate to the production processes being coordinated. Using economic-value units requires flexible and changing, marginal scarcity-based prices for valuation, as well as giving significant autonomy to subordinate units that inevitably then will make the trade-offs in assortment, quantity and timing within planned constraints on values (that is, 'budgets'). Hence, such valuations pose a fundamental challenge to the command economy.

Planning in physical terms, however, leads to 'enormous waste and inefficiency, to production for waste as much as for use' (Grossman 1963, p. 110). There are at least three fundamental sources of this elemental waste: *grossness*, *aggregation*, and *unit of measure.* The need for these arises from the overwhelming complexity of the task of planning for, and directing the operation of, a complex social economy and the necessarily limited information gathering, processing, and dissemination capabilities of any economic agent or agency. However, the emphasis on *gross output* leads to 'input intensiveness', waste, and ignoring cost considerations. *Aggregation* leads to persistent subcategory imbalance in assortment, quality, type, timing, and so on, while *units of measurement* determine suboptimization objectives, distorting implementation decisions, particularly when they are, for material balance reasons, input oriented. Thus each of these is essential for the feasibility of directive central planning, of the command mechanism, yet each loses, or destroys, essential information for the 'proper' (in the eyes of the system directors) implementation of plans, and opens space for creative interpretation of instructions/commands, and hence for 'suboptimization' by implementing units whose interests are not perfectly aligned with those of the centre (Nove 1977). While the command mechanism logically requires unauthorized initiative to be forbidden, and strictly punished when exercised, the size of the task it faces inevitably opens the opportunity, indeed often the need, for such unauthorized initiative. Thus the physical quantity planning required by the command economy to maintain minimal functional 'balance' contains its own antithesis, unleashing forces that undermine the consistency of the plan and the coherence and balancedness of its realization. This fundamental contradiction lies behind most of the critical problems of the command economy in the Soviet Union and the myriad efforts to resolve them within the framework of the command mechanism that comprise the endless waves of reform following victory in the Great Fatherland War of 1941–5.

The 'logic of command' thus imposes a need to restrict autonomy, to restrict the capability of economic units to pursue any other than 'planned' or commanded purposes: economic agents must not have the capability to autonomously acquire and deploy resources for any purposes outside the plan. Comprehensive material balance planning and centralized materials and equipment allocation provide a necessary component, but one that is insufficient unless resources, including human, are denied the capability of autonomous movement and application. Severe restrictions on labour mobility, albeit not as severe as under Stalin, are required, as are comprehensive restrictions on the use of any 'generalized command over goods and services' – that is, money – that might be used to alter their patterns of allocation and use in the economy. The system must be substantially demonetized in order to '... constrict the . . . range of choice in the face of the state's demands' (Grossman 1966, p. 232).

Thus money must be deprived of 'moneyness' and prices must be kept 'passive', as mere accounting and measurement units. According to the logic of the command economy, the availability of money and the prices at which commodities and products are provided should have no essential impact on the allocation of goods and services,

or on the nature and direction of economic/industrial development; all real activity is preordained in the plan and its subsequent implementing commands. The role of money is then to facilitate monitoring of commanded performance through the financial flows it generates. Thus monetary prices do not, and indeed should not, reflect to a substantial degree social goals and priorities; they merely reveal and measure the flow of commanded activity. Producer prices (and most retail prices), wages, prices of foreign currencies, and so on are generally centrally set and controlled, often remaining fixed for long periods of time. Micro-disequilibria naturally abound, while the widely perceived dubious meaningfulness of such prices and the administrative allocation of most producer goods in physical terms combine to sustain the system of detailed production plans and directives in terms of physical indicators – yet another bar to more efficient planning and management.

Finally, an absolutely essential, indeed defining, institution of the command economy is the *physical rationing* of resources and producers' goods. This is where the market is most fully and directly replaced, and where the central authorities have the ability to most directly influence and control the behaviour of subordinate operational units. It implements the centralized mobilization of resources to priorities, the most direct response to crises and challenges. And it most directly denies to subordinates the capability to produce, to develop, in ways outside those authorized in the plan. This makes the coexistence between the command principle and the market mechanism a source of continual conflict, as the market opens unauthorized opportunities to subordinates. In the Soviet Union the command principle, aided by the club of materials rationing, repeatedly pushed back and eliminated the market mechanism when (timidly) introduced in reforms, until the system collapsed in chaos, and the introduction of a full-fledged market economy was begun in 1992 (Schroeder 1979; Aslund 1995). Thus the nature of the command system makes it fundamentally incompatible with real markets, although some market institutions can, and indeed must, be allowed to function both within the non-state sectors and as the interface between them and state economic institutions/sectors.

Inherent Challenges to the Command Economy

As Grossman notes in his seminal article (1963, p. 107), 'The chief persistent systemic problem of a command economy is the finding of the optimal degree of centralization (or decentralization) under given conditions and with reference to given social goals'. The fundamental dilemma is that full centralization poses an insoluble problem, while decentralization abandons the ability to direct, to control development, and to ensure the pursuit of social goals and priorities. With regard to the pure planning problem, a large body of theoretical literature arose in the late 1960s, and continued into the 1980s, on the problem of decentralizing the planning process to make its informational and computational burden manageable (Eckstein 1971; Bornstein 1973). But the problem is far greater, and less studied, with respect to implementation; rational planning is swamped by the struggle to maintain elemental coordination.

Decentralization Versus Priority

Looked at through the prism of relative advantages, operational decentralization shortens 'lines of communication', increasing flexibility, adaptation and responsiveness to a changing environment through local initiative and innovation, and vastly simplifying the decision problem of economic agents. But it does so at the cost of weakening or losing the 'advantages of centralization', including enforcement of regime values, capability for large-scale resource mobilization, concentration of scarce talent in central decision-making organs, and the maintenance of macro-balance. In particular, decentralization compromises the ability of the centre to directly manage the development and structure of the economy and to force the achievement of critical priorities regardless of cost. Furthermore, decentralization requires the introduction

of the alternative coordination mechanism to insure tolerable micro-balance – the market – as decentralization undercuts the ability to directly coordinate, to balance from above. Thus, to prevent catastrophic imbalance, a more active money with economically flexible market prices must be allowed to function in a decentralized system.

The impossibility of planning and commanding the performance of all economic agents in full operational detail, however, forces some decentralization. This creates a chronic threat to balance which is thus a continuous argument for (re) centralization of planning and materials allocation. Furthermore, a partial decentralization of planning and management in a command economy may do more harm than good; it may impair balance without yielding sufficient benefit. Yet a complete decentralization, in the sense of a virtually full devolution of the major production decisions to the firm level, would be disastrous from the standpoint of balance, unless the price structure were properly altered to provide proper signals to firms *and* suitable behavioural rules were prescribed, that is, unless a market mechanism were introduced. Thus the logic of command predicts a 'treadmill of reforms' (Schroeder 1979), an array of countervailing strengthenings of the oversight and control organs (in particular, the Party), and enhancements of their role in the economy, accompanying moves towards decentralization in the state sector. It also explains the Soviet institutional arrangement of inter-firm contracts as a decentralized implementation device. These are required to specify details of interaction within planned categories, and establish observable, and hence legally enforceable, commitments to planned implementation, constraining the autonomy necessarily granted through the minimal decentralization. And it explains the logic of the continuing restraints on the use of money and the continuing efforts at effective price control to keep the autonomy of agents restricted to the minimum necessary for the continued functioning of the less-than-absolute command economy.

Even limited decentralization requires that money be used in the command sector (as well as in the household sector), but its role as a bearer of options and as the means of pecuniary calculation for decision-making is necessarily limited and deliberately subordinated to the planners' will and the administrators' power. Banks and the treasury accommodate the money needs of production, ensuring a soft budget constraint for the individual firm. At the same time, the 'moneyness' of money at the firm level is low, hemmed in as it is by administrative constraints and impediments, including the rationing of nearly all producer goods, and by the widespread 'seller's market' (shortages of goods and absolute lack of buyers' alternatives). This monetary ease, together with the sellers' market, plays an important role in ensuring individual workers' job security at the firm level and full employment in the large, while keeping the firm largely insensitive to money costs and/or benefits.

Within the command sector, money and prices have a necessary role in determining terms of alternate resource uses *only within* planned/commanded categories, and money has the role of limiting total claims to resources in areas, or at a level of detail, beyond the reach of plan directives. This requires 'businesslike management' within the firm – *khozraschet*, which is a 'set of behavioral rules that is supposed to govern the actions of Soviet managers beyond their primary responsibility, the fulfillment of output targets'. It pushes the firm toward 'technical efficiency' and limitation of 'claims on society's resources for productive use. . . . *khozraschet* is a system that is well devised to control the behavior of managers in a command economy where a certain amount of devolution of power to them is inevitable, and where, further, managers' goals and values do not necessarily coincide with the official ones' (Grossman 1963, p. 117). Thus money also has the role of facilitating the monitoring of performance in the command sector.

While administrative orders are the rule in a command economy, backed up by greater or lesser degree of state coercion (depending on country and period), any decentralization of implementation naturally relies on monetary ('material') incentives to elicit desired individual compliance and performance. Compounding the incentive problems arising from differences in information and interests between central authorities and implementing

agents is the fact that the physical and other indicators to which the material incentives are linked may often be poor measures of social benefit (as seen by the leadership). Furthermore, resort to such rewards widens the distribution of official earnings and raises questions of permissible limits of income inequality. Yet there may be little choice in that the state must in effect compete with the much higher incomes from the *second economy.* Indeed, the Soviet Union during War Communism, Cuba in the 1960s, and the People's Republic of China during some periods before Mao's death in 1976 tended to downgrade material incentives in favour of normative controls, but never did quite abolish them.

The behaviour of the Soviet-type firm has been much studied (Granick 1954; Berliner 1957; Nove 1977; Freris 1984). Because its directives and the corresponding managerial incentives stress physical output, produced or shipped, and thanks to its low sensitivity to cost and the ambient sellers' market, the firm often sacrifices product cost, quality, variety, innovation and ancillary services to its customers to sheer product quantity. By the logic of command and the requirements of plan manageability, firms operate in an environment with sole suppliers and assigned users, reducing complexity by eliminating 'wasteful' redundancy in production and distribution. Thus firms in a command economy are largely insulated from any product competition, both from the outside world and from other domestic firms, thanks to the climate of administrative controls and the prevalent excess demand for their output. Difficulties with supply, frequent revision of its plans, interference by Party and other authorities, and other systemic problems also stand in the way of its more efficient and effective operation. Indeed, to function at all, the firm's management is frequently forced to break rules and even resort to criminally punishable acts.

This compounds a further critical challenge posed by necessary decentralization – the conflict between the will, purposes, incentives and priorities of the higher authorities and those at lower levels, particularly of the firms and their managements. Even the best-motivated managers, following all official rules and incentives, will sometimes fail to replicate the decisions that their superiors would have made had they been in a position to make them. This problem is aggravated by the inevitable ambiguity, incompleteness and inconsistency of those rules, incentives and the information available on the spot. Only binding physical constraints and observable outcomes can be systematically enforced, making 'centralized materials allocation the most powerful weapon at the disposal of the central authorities' (Grossman 1963, p. 118). Thus, where material inputs are less determinate of a unit's activities, this information and incentive problem is greater, and the defiance of central will relatively more widespread and successful. This observation explains the non-viability of any reform that fails to fundamentally alter the materials allocation system.

Under-Planned, Ill-Commanded Sectors

A major challenge to the command economy also arises from the existence of sectors outside, or only partially affected by, the command principle. In the Soviet Union these included most of agriculture, much of housing, the household sector and some consumer goods and services. 'Markets' were allowed to function for the distribution of final consumer goods and services, including agricultural produce, for much of the activity of the 'collective' sector in agriculture and for household labour supply. For transactions with 'personal property' within the household and collective sectors, money was active and agents responded to market prices, while in the quasi-markets interfacing with the state sector – for example, labour and consumer goods – money was relatively active but prices remained largely controlled and non-market. These are sectors where information on needs/preferences and capabilities proved too difficult to acquire reliably in real time for acceptable allocation and balance to be commanded, and so at least one side of a market was allowed to function with an active money. Here, the command mechanism proves too crude and clumsy, and hence politically counterproductive, to be used outside of pressing emergencies. Indeed, this might be considered a lesson of War

Communism, the first experience with a command economy in Soviet Russia, 1918–21.

In view of the theoretical incompatibility of command and market, how could these 'market' sectors be successfully grafted on to the command mechanism? An explanation (Grossman 1963) rests on the trade-offs between the authorities' limited capabilities, the complexity of those sectors, and their centrality to regime priorities. A sector which provides significant inputs to physical planning and plan fulfilment, where the unpredictability in the flow of goods is unacceptable, cannot be left to the market without seriously undermining command. However, if a 'market' sector can be treated as a residual for purposes of materials planning and allocation, a buffer for planning, then its coexistence is acceptable. Further, if its operation is characterized by rapid change and complexities rather outside the core interests of the regime, if without disrupting the industrial core greater incentives and risk can be placed on those peripheral agents, and if non-market constraints can force the desired market response from it, then the centre will want to separate that sector from the command sphere, lowering its coordination burden by shifting it to the market.

These considerations were indeed active in the case of those sectors 'left to the market' in the Soviet Union: consumer goods retailing, the acquisition of labour services, the support of households in the countryside through a private agricultural sector, and a few peripheral and interstitial activities. Indeed, any attempt to truly 'marketize' any other sectors or activities in the command economy is doomed to fail unless the loss of fervour, of the sense of mission and urgency, leads to abandonment of the command mechanism. Yet even the existence of these limited market sectors, providing an outlet for incentive earnings and diverted resources, exerts a continuing corrosive pressure on the command economy and its control mechanisms.

The Cancer of 'Money'

A truly monumental challenge to the command economy lies in the role of money in any *less-than-absolute* command economy. As the complete centralization of decisions in the production sector (let alone in the household sector) is an impossibility, something must be left to local initiative and dispersed decision making. Thus *khozraschet* is a logical necessity, '... an unfriendly bridgehead that threatens to seize ground whenever the planner fails or defaults' (Grossman 1966, p. 228). With the inevitable devolution of some decision making to firms and households, money acquires a necessary and critical role in the command economy, going well beyond that consistent with the logic of command. That role arises from the need to economize in making decentralized decisions, and as a medium of exchange and store of value in the decentralized interactions that relate to all decisions. In acquiring this role, this 'moneyness', it allows accumulations of power outside the control of the regime. Money is a 'bearer of options' whose power and influence must be restrained if the command mechanism is to operate properly – to determine priorities and to insure maximal commitment to their achievement. As Grossman (1962, p. 214) noted, 'Money is a form of social power that may lead resources astray and is subject to only imperfect control by political authority.'

Thus the power of money has to be curbed in a command economy by limiting balances available to households and firms, by compartmentalizing money into cash and 'firm' circuits, and by erecting barriers and limits to the use of 'monies' in each category, although that undercuts the effectiveness of attempted decentralizations. Liquidity, 'moneyness', is constrained by the institutional structures and by all the characteristics and conditions of the 'sellers' market', rendering 'money' the only non-scarce commodity, in unusable excess to the extent the command mechanism is effective. Monetary policy in the properly functioning command economy reduces to limiting the volume of cash in the economy ('macro-monetary' control) through wage fund restrictions and cash control absorption plans of the retail sector, and the allocation of firm balances in restricted categories ('micro-financial' control) in just sufficient quantity to support the implementation of the plan, with confiscation of

excess funds to prevent unauthorized activity by the firm (Garvy 1977).

Similarly, the price system, expressed in terms of that money, must also be mobilized to the purpose of control. The inflexible, administratively segmented, average cost-based prices in command economies are a logical necessity of command- and haste-based shortage. For all the problems they cause, all the unintended consequences and distortions in the behaviour of subordinates, such prices help to keep money largely passive, at least in the core state sectors, and allow both money and prices to remain instruments, rather than disrupters, of command. More than being ideologically justified, such prices are a response to the pragmatic and pressing requirements of running a shortage economy with a rapidly developing system of centralized direction of enterprises and of materials allocation.

Money, however, is not so easily contained. Once in unobserved hands, it exercises its 'command over goods and services' without reference to plans, commands or regime priorities. Hence, given any discretion, in any sphere of activity not directly monitored agents will naturally use money in ways they find desirable, placing new demands on a physical system otherwise tautly planned and characterized by general scarcity. This is facilitated by the existence of agents and spheres of activity outside the command system, providing 'legitimate' sources and uses for monies, however acquired or disposed. And the possibility of acquiring money provides incentives for unauthorized activities, incentives to undertake unplanned interactions and reallocations. An active money vastly expands the sphere of discretion of 'subordinate' agents beyond any authorized by a decentralizing reform, and calls for severe administrative restrictions, a reduction to passivity, if it is not to disrupt the planned activities and discretion of the central authorities.

Yet attempts to administratively constrain the influence, the 'corruptive' power, of money become increasingly futile once the 'genie' has been 'let out of the bottle'. Even limited decentralizing reform, allowing money to influence some (subcategory) production and allocation decisions, inevitably lets loose more liquidity, more of a command over goods and services, than desired. This arises from a multitude of factors: errors in both physical and financial plans, inherent incompleteness of plans and commands due to limited information and time and to the necessity of aggregation, changing circumstances and shocks to the economy, mistakes in implementation and in responding to shocks, the irregularity and disruptions in the materials allocation system, the behavioural response of even the most enthusiastic and best-intentioned agents to these problems, and so on. All of these can lead to an unexpected lack of funds for doing what was commanded (if only implicitly), and hence disruption of commanded performance, unless additional liquidity is provided.

Thus monetary policy in a command economy, once money is allowed any room for activity, must be accommodating; a lack of funds can never be allowed to disrupt planned performance, just as an excess of funds cannot be allowed to facilitate unplanned or unauthorized activity. Thus the role, the influence, of money has a natural, inexorable tendency to grow: insufficient funds become an immediate problem generating new money through credits or additional allocations, while unused funds tend to stay hidden until ferreted out by inspection or accidental discovery. And as it grows, so does the challenge to the command principle. An increasing number of agents, in both the state and non-state sectors, has a growing ability to access resources, to divert them in the name, if not always the interest, of implementing decentralized plans, and thus to challenge the priorities of the political authorities. This growing challenge becomes a cancer in the system, a growth that undermines its health and feeds tendencies destructive of the priorities of the regime and its rulers.

The 'Second Economy'

As the command economy matures, as the messianic fervour with which it was imposed wanes and the use of extraordinary force diminishes in ensuring compliance with commands, these challenges to command metastasize into a competing yet

symbiotically attached and dependent economic system: the *second economy.* This name highlights the distinction of this sphere of economic activity from the officially sanctioned, 'first', command economy. It is thus defined as 'all production and exchange activity that meets at least one of two criteria: (a) being directly for private gain; (b) being in some significant respect in knowing contravention of existing law' (Grossman 1977, p. 25).

In the Soviet Union, attempts to strengthen 'material incentives' and activate 'the profit motive' in order to increase the effectiveness and technical efficiency of the implementation of central plans and directives and to stimulate technological progress and innovation, and the growing monetization of the agricultural sector, opened the door to massive expansion of money supply and eroded the barriers between the currency and the enterprise bank account monetary circuits. Collective farms and their subsidiary enterprises, owners of 'small means of transport', vodka manufacture and distribution, and the Caucasus republics (Georgia in particular) proved particularly rich sources of illicit (from the system's perspective) monetization and private 'entrepreneurial' activity. This both raised the spectre of inflation and opened the door to vastly increased opportunities for manipulation by self-interested subordinates in the command sector. Thus the use of 'economic levers' greatly increased the opportunity for and incidence of bribery, corruption, speculation, and even 'honest' private labour.

While the fundamental cause of the appearance and growth of the 'second economy' undoubtedly lies in the congenital institutional weaknesses of the command economy discussed above, there are a number of proximate sources that make it unsurprising. These include extensive price control, with consequent scarcity and misallocation, high taxes on non-state activities/incomes, prohibitions of private activity, unmet individual consumption needs, poorly protected impersonal (state) property, the personal power of bureaucrats and 'gatekeepers', and other historical factors, including the end of terror. These provide both motives and opportunities for officially illicit activity and for the authorities to overlook that activity. With the ageing of command and the decay in enthusiasm of its agents, the growth of such a second economy appears natural.

Thus growing 'monetization', the existence of ready and waiting market sectors, and the decline in the use of violent instruments of enforcement lead to a growing sphere and importance of activities outside the purview of 'planning' and 'command'. These market-mediated activities are at times supportive, helping to achieve tolerable micro-balance in the increasingly complex economy, but often are in violation of planned implementation and regime values. Private interests, necessarily allowed some leeway, grow in significance, increasingly seizing ground from command. In the Soviet Union, the private agricultural sector, initially permitted only to secure survival of the peasantry under the extractive pressure of rapid industrialization, and the consumers' personal services sectors provided the basis for a ubiquitous, if still systemically marginal, second economy.

But then even the core industrial sectors under the command mechanism find their managers and activities increasingly influenced by this illegitimate, shadow market, system, as managers are often forced to break rules and undertake illegal acts in order to do their job. Such acts, together with ubiquitous and protean illegal activity on private account, add up to a large underground economy characteristic of every command economy, which together with legal private activity (allowed in varying degree in different countries) both supports and supplements the 'first economy' and is inimical to it. While the second economy significantly adds to the supply of goods and services, especially for consumption, it also redistributes private income and wealth, contributes to the widespread official corruption, and generally criminalizes the population. Virtually every area of economic life is touched upon, and often entangled with, 'second economy' activities, while legal private activity naturally opens a loophole for illegal trading and entrepreneurship, generally below the purview of the authorities. And it goes hand in hand with the extension of corruption, ensuring that it remains outside of official notice.

Those 'violations' of legality within the command sector, a 'shadow economy', build informal

inter-enterprise relations which are generally beneficial to the operation of state enterprises. They work to substantially correct the allocative failures of the command mechanism, improving firm performance and hence benefiting its management, and also provide lucrative opportunities for managers to directly benefit through the activization of barter, personal connections, and bribery. However, they also spawn further distortions in economic behaviour, as managers seek to generate access to cash, the life blood of the 'second economy', to extract rents, and to hide their activity from supervisory and statistical organs.

Thus the second economy plays a dual and contradictory role in the command economic system. First, it addresses a number of the problems of coordination and balance endemic to the command mechanism, reallocating both producers' and consumers' goods, facilitating plan fulfilment and the use of financial incentives, and generating new incomes and 'politically safe' outlets for private initiative. Hence it becomes important for enhancing consumer welfare, for production stability, and even for social stability. The 'second economy', and in particular its 'shadow' side, plays an essential role in the first economy as a 'pressure valve', a release 'fixing command' by maintaining micro-balance and covering 'holes' in economic life left by the mistakes or oversight of the planners and central managers. And this role becomes increasingly important as the economy grows more complex and diversified, and hence becomes less susceptible to conscious oversight and direction.

As the central authorities struggle with their loss of control, searching for a solution through reform, decentralization and recentralization, monetization and administrative restriction, agents in the economy take advantage of gaps in control, of the autonomy and discretion offered by growing liquidity of the quasi-money in the system, to deal with problems of coordination and balance, inconsistency of plans and commands, and ubiquitous shortages and scarcities. Of course they operate in light of their own partial information, and in their own (private as well as official) interests, but in so doing save the system from collapsing under its own weight and rigidity (Powell 1977). Thus the second/shadow economy provides a spontaneous surrogate economic reform that imparts a necessary modicum of flexibility, adaptability and responsiveness to a formal set-up that is too often paralyzing in its rigidity, slowness, and inefficiency. In doing so, the second economy also provides a valuable stabilizing influence on society and the polity, making life livable and the system humanly manipulable and responsive to private inducement. It makes everyone complicit in the way things work, equally 'guilty' before state and society, while providing an almost legitimate, and not politically dangerous or directly destructive, outlet for individual initiative and entrepreneurship. Finally, it relieves inflationary pressures (a 'monetary overhang') resulting from the command economy's necessary combination of monetary looseness and pricing rigidity.

Despite this positive functional role, the second economy also has a less positive *systemic* impact. It mocks the pretense of social direction and control, subverts its egalitarian impulse by accentuating differences in access and income, and gives the lie to the pretense of a 'new' ideologically correct ('Soviet') man. Its very existence and usefulness thus subvert the ideology of the regime, and it works against and undercuts regime priorities by exposing the incompetence and incapacity of the authorities. Its provision of alternatives weakens the 'plan, production, and labour discipline' so essential to the proper operation of the command mechanism. Indeed, it attacks the core of the command mechanism as it '... elevates the power of money in society to rival that of the dictatorship itself, rendering the regimes implements of rule less effective and less certain' (Grossman 1977, p. 36). In particular, it corrupts officialdom and distorts prices, adding a (positive or negative) 'second economy margin', both 'in kind' and in money, breaking prices as an effective instrument of control. This weakens monetized incentives for state activities by providing competing, and often better, alternatives to them. Hence the second economy, and in particular the 'shadow economy' in the state sector, completes the cancerous development of agent autonomy, of the ability to work outside the plan and its

subsequent commands, by providing viable alternatives to the plan.

Other dysfunctional impacts, undermining the operation of the command system, arise from its diverting of resources and products to unplanned sectors and activities, including diversion from development/investment priorities to consumers. This naturally generates undesirable (from a system perspective) redistribution of incomes, although recipients, including many high-placed officials, find it very desirable. Indeed, it is further disruptive of command by creating a 'two-tiered' system of prices and incomes, of consumer goods and labour markets. One tier is comprised of the low-priced, scarcity-ridden quasi-markets of the 'less-than-absolute' command economy, where the unenterprising, the overly scrupulous, and the 'slow' can survive. The other tier consists of real, albeit highly distorted, markets in the generally high-priced, risky but well-endowed second economy where the enterprising, entrepreneurial, and criminal can thrive. In this high tier, substantial incomes are generated and allocated, although they largely accrue to corrupt officials and 'gatekeepers' of scarce materials or permissions who can extract rather phenomenal 'rents'. The inequities this generates further undermine the legitimacy of the regime and generate potentially explosive social pressures, only partially relieved by the second economy's 'pressure valve' aspects.

Finally, it is worth noting that the second/shadow economy, through its activity outside of the officially measured sphere, seriously distorts statistical data and the information available to planners and allocators in the official economy, and, due to its illegality, also hides necessary information from other agents in the shadow economy. This aggravates the economic problems that spawn 'second/shadow economy' activities, deepening the contradictions between the centre and decentralized agents, and further corroding the institutional structures of the command economy.

Performance and Fate

Command economies have been instrumental in radically transforming societies more or less according to their drafters' intents, in mobilizing resources for rapid industrialization and modernization, at times on a vast scale, and in rapidly amassing industrial power and military strength. Indeed, they have shown themselves highly effective in rapidly implementing large-scale projects and achieving overriding social goals, albeit at great cost. It is this effectiveness, when cost is no object, which explains why the command principle is resorted to in times of emergency and war. Hence in the Soviet Union command facilitated defence during, and rapid recovery and rebuilding of the Stalinist economy after, the massive trauma of the Great Fatherland War. Economic growth has been especially marked (though not unparalleled by market economies) where large amounts of unemployed and underemployed labour and rich natural resources could be mobilized and combined with existing (advanced, Western) technology, and where the public's material improvement could be restrained, or even seriously depressed, under strong political control. As these possibilities waned, and as the economies grew in size and complexity and thus became less amenable to centralized administrative management, rates of growth declined sharply. At the same time, the shortcomings of the command mechanism in adapting production to demand and its changes – providing consumer welfare, effecting innovation, serving export markets – became more apparent and less tolerable. This led to much discussion and repeated attempts at controlled institutional reform, at decentralizing and stimulating subordinate initiative without sacrificing ultimate control.

Some actual reforms in the externally imposed command economies of eastern Europe went so far as to introduce or extend the market mechanism to such a degree that one could no longer regard the system as a Soviet-type command economy, even if, before the 1990s, one could not speak of it as a full-fledged market economy either. Yugoslavia since the early 1950s, Hungary since 1968 and especially in the 1980s, and post-Mao China are the most important cases in point. Other actual reforms were of a minor or 'within-system' nature, aiming to decentralize certain types of decisions while eschewing the market

mechanism and retaining the hierarchical form of organization and the command principle. In the hope of stimulating efficiency to revive growth rates, the decentralizing measures were accompanied by a number of other 'reforms' relating to organizational structure: prices (still controlled), incentives, indicators, materials rationing, and so on. The Soviet reforms of 1965, and those in the 1970s and 1980s prior to *perestroika*, were of that kind; many similar ones took place in other Communist countries after the mid-1950s and prior to the overthrow of Communism in 1989. On the whole, such reforms had little success in addressing the problems of the command economy. Bureaucratic and political obstacles apart, the attempt to decentralize economic decisions without bringing in a market mechanism almost inevitably leads to economic difficulties. The beneficiaries of devolution of decision-making lack the necessary information to produce just what the economy requires or to invest to meet prospective needs, and the coordination of plan-subsequent command is lost. Moreover, they may apply the additional power at their disposal to advance particularist causes or to divert resources into illegal channels. Microeconomic disequilibria mount, and soon superior authorities step in to recentralize on a case-by-case basis and the reform withers away (Grossman 1963; Wiles 1962, ch. 7; Kontorovich 1988).

This failure of reform reflects the inherent contradictions of the command economy framed in the irreconcilable conflict between 'command' and 'money' discussed above (Ericson 2005). The Soviet command economy, driven by the urgent need for and haste in industrialization and military development, initially relegated the influence of money and the market to the margins of the system, where they handled areas and activities in which command had been revealed as counterproductive during War Communism. That system, the 'less than absolute command economy', substantially industrialized, triumphed in the Great Fatherland War, and recovered to an almost perfect replica of its pre-war self by 1950. But by then the strains of its inherent inflexibility and the bounded rationality of the system's planners and managers began telling on continuing growth and the development of the economy. With economic growth came increasing complexity and growing intractability of the central planning and economic management problem. Some decentralization became essential, and increasingly so as time passed, opening the door once again to the rise of money as a significant influence on the operation and development of the economy. And that influence was only enhanced by the ageing and mellowing of the system. With the passing of 'terror' as an effective incentive mechanism, the stabilization of personnel and the regularization of procedures, it became ever harder to control agent behaviour, to contain the distractions of money and the self-interests it mobilized, and to uncover the rents that well-placed agents were able to extract, thus aggravating the inherent agency problems of the command economy.

The remaining years of the Soviet system thus witnessed an epic struggle, barely perceptible at first, but increasingly evident as reforms, decentralizations, reorganizations and recentralizations cycled around each other in the search for a solution to the increasingly evident and destructive malperformance and waste, and aggravating behavioural distortions in response thereto, generated by the struggle between the 'command principle' and the weak, but inexorably emerging, 'market'. Initially reflected in the dysfunctions of the marginal and quasi-markets of the command economy, and in the struggle to harness a 'passive' money to the purposes of command, the role of money grew along the 'treadmill of reforms' into the rival, if still largely subordinate and complementary, 'second economy', and in particular its 'shadow' component, on which the 'command principle' increasingly came to depend for its effectiveness. As long as the Soviet system remained a 'command economy', commands had to have last word, and money remained largely relegated to the sidelines, exercising its influence within the quasi-monetized instruments ('economic levers') of the command mechanism and the distorted markets of the second economy.

This inherent conflict, played out over Soviet history, revolves around a number of fundamental

dualities, elemental oppositions which characterize these primary forces. The 'command principle' derives most basically from the urge, the will to control, to 'rationally' determine and direct the future, exercised by a 'gnostic' elite, immanent in the Party. It knows what needs to be done, by whom and how, and can tolerate no dissent or deviation. Juxtaposed to this 'Will of Society' stand the millions of independent 'wills', desires and objectives, anarchically coordinated through 'the market', whenever that set of institutions broke through the barriers and limits placed by 'command'. This provides the foundation for the eternal struggle between 'central priorities and control' and 'agent incentives and capabilities'.

This opposition is severely aggravated by urgency, by 'virtuous haste', in the pursuit of overriding social goals and central objectives. For the mobilization for, and focus of resources on, these priorities trample on the information, capabilities and goals of individual and organizational agents which must perforce implement that mobilization, implement those priorities. 'Effectiveness' in the pursuit of social objectives becomes opposed to 'efficiency' in the attainment of any objectives, denies trade-offs based on local information and incentives, and hence blocks flexibility in response to changing circumstances. Indeed, the single-minded pursuit of overriding objectives, of absolute priorities, naturally disrupts the fine coordination, the requirements of 'balance', necessary to consistently and efficiently pursue any objectives.

Throughout the history of the Soviet Union, the needs of centralization, given Soviet social goals, stood in fateful opposition to the necessity to decentralize in order to keep the system tolerably functioning. The latter necessity spawned repeated (partial) remonetizations and a 'second economy' that both shored up the operational foundations of the 'first economy' and undermined its long-term viability, corroding its ideological and systemic foundations. Money so unleashed intensified the dysfunctions and contradictions of the 'command economy', spurring further repeated 'reforms' and 'experiments' that merely further aggravated the inconsistencies, the 'oppositions' in the system, until the central leadership, largely unintentionally and out of ignorance, destroyed the 'command economy' in the radical systemic and economic 'restructurings' beginning with *perestroika* in 1987.

See Also

- ▶ Agency Problems
- ▶ Decentralization
- ▶ Informal Economy
- ▶ Second Economy (Unofficial Economy)
- ▶ Soft Budget Constraint
- ▶ Soviet Economic Reform
- ▶ Soviet Growth Record

Bibliography

Aslund, A. 1995. *How Russia became a market economy*. Washington, DC: Brookings Institution.

Berliner, J. 1957. *Factory and manager in the USSR*. Cambridge, MA: Harvard University Press.

Bornstein, M., ed. 1973. *Plan and market*. New Haven: Yale University Press.

Eckstein, A., ed. 1971. *Comparison of economic systems: Theoretical and methodological approaches*. Berkeley: University of California Press.

Ericson, R. 1983. A difficulty with the 'command' allocation mechanism. *Journal of Economic Theory* 31: 1–26.

Ericson, R. 1991. The classical Soviet-type economy: Nature of the system and implications for reform. *Journal of Economic Perspectives* 5 (4): 11–27.

Ericson, R. 2005. Command vs. shadow: The conflicted soul of the Soviet economy. *Comparative Economic Systems* 47 (4): 1–27.

Freris, A. 1984. *The Soviet industrial enterprise*. New York: St Martin's Press.

Garvy, G. 1977. *Money, financial flows and credit in the Soviet Union*. New York: Ballinger Publishing Company.

Granick, D. 1954. *Management of the industrial firm in the USSR*. New York: Columbia University Press.

Grossman, G. 1962. The structure and organization of the Soviet economy. *Slavic Review* 21: 203–222.

Grossman, G. 1963. Notes for a theory of the command economy. *Soviet Studies* 15 (2): 101–123.

Grossman, G. 1966. Gold and the sword: Money in the Soviet command economy. In *Industrialization in two systems*, ed. H. Rosovsky. New York: Wiley.

Grossman, G. 1977. The 'second economy' of the USSR. *Problems of Communism* 26 (5): 25–40.

Grossman, G. 2000. Central planning and transition in the American desert: Latter-day saints in present day sight. Available as a preprocessed paper at http://www.econ.berkeley.edu/~grossman/mormons.pdf. Accessed 23 June 2006.
Higgs, R. 1992. Wartime prosperity? A reassessment of the U.S. economy in the 1940s. *Journal of Economic History* 52: 41–60.
Kontorovich, V. 1988. Lessons of the 1965 Soviet economic reform. *Soviet Studies* 40: 308–316.
La Lone, M., and D. La Lone. 1987. The Inka state in the southern highlands: State administration and production enclaves. *Ethnohistory* 34: 47–62.
Nove, A. 1977. *The Soviet economic system*. London: Allen and Unwin.
Powell, R. 1977. Plan execution and the workability of soviet planning. *Journal of Comparative Economics* 1: 57–76.
Raupach, H. 1966. Zur Entstehung des Begriffes Zentralverwaltungs-wirtschaft. *Jahrbuch für Sozialwissenschaft* 17 (1): 86–101.
Schroeder, G. 1979. The Soviet economy on a treadmill of reforms. In *Soviet economy in a time of change*, ed. Joint Economic Committee, US Congress, vol. 1. Washington, DC: Government Printing Office.
Wiles, P. 1962. *The political economy of communism*. Cambridge, MA: Harvard University Press.
Zaleski, E. 1968. *Stalinist planning for economic development*. Durham: Duke University Press.

Commodity Fetishism

Andrew Levine

Abstract

An analysis of Marx's notion of 'commodity fetishism' – as a theory of the necessary (systemically induced) misperception of underlying production relations by participants in market exchanges. The appeal of the notion to the two main opposing tendencies of mid- and late 20th-century Marxism – Marxist humanism and structuralist Marxism – is discussed. Reasons are proposed to account for a recent decline of interest in the phenomenon among both economists and philosophers. It is suggested, however, that the concept remains viable.

Keywords

Althusser, L.; Capitalism; Capitalist social relations; Commodity fetishism; Exploitation; Feuerbach, L.; Invisible hand; Labour power; Labour theory of value; Laws of motion of capitalism; Lukacs, G.; Market relations; Marx, K.; Marxist humanism; Structural marxism; Surplus value; Value controversy

JEL Classifications

B1

Since Plato, philosophy and then science have assumed first, that there is often a difference between appearance and reality; and, then, that it is sometimes possible to grasp what really is the case by investigating how things appear. Marx's account of commodity fetishism, a crucial step in his account of the capitalist mode of production, implements these assumptions explicitly. It describes how exchange relations appear to economic agents, where the appearance belies the reality at the same time that it provides cognitive access to it.

Market exchanges occur in all modes of production capable of sustaining an economic surplus. In capitalism, the process is generalized – not just in the sense that markets structure economic life but also, more importantly, because everything is commodified that can be. Universal commodification is the result of a protracted process that is definitively launched once labour – or, more precisely, labour power (labour time, adjusted for differences in intensity) – is commodified. The commodification of labour power is pivotal because this commodity is the sole source of value and therefore, ultimately, of wages, profits and rents. The generation and distribution of surplus value, of what is produced in excess of what is needed to reproduce the labour power expended in production processes, is the invisible underlying reality upon which perceptions of exchange relations depend. To persons engaged in buying and selling labour power, what appears is just that, as in any other exchange, individuals aim to do as well for themselves as they can, given

their resources, their preferences, and the production technologies available to them. But what is really going on is a struggle over the distribution of the economic surplus at the point of production. That reality is opaque. Economic agents are therefore governed by the appearance of rational economic agents maximizing payoffs to themselves. In his account of commodity fetishism, Marx shows how this inevitable misperception helps to reproduce and sustain the underlying reality.

When Marx expressly addresses this phenomenon at the conclusion of the opening chapter of the first volume of *Capital* (1867), the economic agents he describes are property-holding individuals. Thus it is not exactly capitalism that he aims to model, but 'simple commodity production', an ahistorical idealization. However, the cogency of his account is unaffected as his analysis becomes more historical and concrete – to the point that the direct producers are, as in full-fledged capitalism, a property less proletariat with nothing to exchange except, of course, their own labour power. Commodity fetishism is therefore a general and pervasive fact wherever capitalist social relations hold sway. Thus the term denotes a systemic opacity at the level of appearance that helps to hold economic agents in thrall by masking the exploitation of labour. Because this misperception sustains the exploitation that engenders it, revolutionaries intent on overthrowing capitalism must tear away the veil of illusion by revealing the exploitation of workers that exchange relations conceal.

Marx does not directly address *how* commodity fetishism comes into being or how it is sustained. But he does provide fragments of an explanation when he focuses on the atomizing effects of market relations. All resource allocation mechanisms are social in the sense that they bring together a host of disparate and heterogeneous economic activities. However, where the commodity form prevails, the social character of market transactions is apparent only after goods and services are produced. The workers know that the corn they consume is produced by farmers, and the farmers know that the tools they use in growing corn are made by workers. Everyone also knows that, without food, workers would not be able to make the tools farmers use; and that, without tools, farmers would not be able to grow food for the workers. It is therefore evident in retrospect that workers and farmers are engaged in a collective endeavour. But it is not similarly evident prospectively. From that vantage point, it seems only that farmers and workers – and also the capitalists who provide them with means of production – are making individual choices aimed at bringing about the best outcomes for themselves, given the constraints they face. Even if they believe that these essentially egoistic activities are somehow socially beneficial, they can justify this belief by appealing to the workings of an 'invisible hand'. Because there is no visible hand that directs the process, the terms of interaction appear *as if* they are forces of nature to which individuals must accommodate. Thus market relations appear as infrangible constraints that human beings are obliged to operate within, not as social constructions that human beings can change. In terms that Kant introduced and that Marx, following Hegel, effectively assumed, freedom (autonomy) is then forfeit. Wills are heteronomously determined, governed by laws of an (apparently) impersonal *other* (the market system itself). To be free, we must therefore take control of the aggregation mechanism we have concocted. We do so by putting reason in command – not just at the individual level of the rational economic agent, but at the societal level as well.

What Marx says about commodity fetishism is concise and intriguing. For these reasons, and because it summarizes the very abstract analysis of the commodity form with which *Capital* begins, his account of the phenomenon has always been well known. 'Commodity fetishism' is one of those terms that everyone associates with Marx. But, even in what remains of Marxist circles, the basic tenets of Marx's account have faded from ongoing discussions. A number of factors have contributed to this turn of events: among them, the legacy of the so-called 'value controversy' of the 1970s; the efforts of mathematical economists in the 1970s and 1980s to put the categories of Marxist economic analysis on a sound, analytical footing; and attempts by analytical philosophers, working on Marxist themes, to reconstruct and, when possible, defend core Marxist positions. The

conclusion that has emerged is that, *pace* Marx, there is nothing special about the commodification of labour power and therefore that the theory of surplus value cannot be sustained in the way that Marx believed. Nowadays, it is only the most doctrinaire Marxists who uphold the labour theory of value, the basis for Marx's account of commodity fetishism. This fact along with the decline of political movements that identify with the Marxist tradition and, its inevitable consequence, waning interest in Marx's work itself, has, for the time being, made commodity fetishism a matter of concern mainly to historians of economic thought.

Not long ago, the situation was quite the opposite. From roughly the 1950s through the 1970s, commodity fetishism played a central role in the two most important and innovative tendencies in Marxist theory: Marxist humanism and structuralist Marxism. These were opposing tendencies, politically and substantively. But they converged on according commodity fetishism centre stage.

Marxist humanists sought to de-Stalinize Marxism by recovering its Left Hegelian roots. This meant reading Marx's work through the prism of his early writings, before he broke with his 'erstwhile philosophical conscience', as he proclaimed in 1845 in *The German Ideology.* For the Left Hegelians, Ludwig Feuerbach's philosophical anthropology, elaborated in *The Essence of Christianity* (1841), was fundamental. There Feuerbach 'inverted' the theological dogma that 'God makes Man' by showing how the God idea is an 'objectification' of essential human traits. Lacking materiality, God is purely an objectification, an 'alienated' expression of the human essence. In taking consciousness of this fact, one recovers essential humanity and becomes emancipated from the thrall of its systemic misrepresentation. In the *Paris Manuscripts* (1844), Marx applied the Feuerbachian programme to objects of labour; 'objectifications' too of essential humanity, but also material things and therefore not objectifications only. Feuerbach arrived at his conclusions by 'interpreting' the theology of Right Hegelian theologians. His working hypothesis was that they had gotten the concept of God right, but that they radically misconstrued what the concept *means*. In the *Paris Manuscripts* Marx treated (Smithian) political economy the same way. He assumed that it correctly describes 'economic facts'. The task, then, was to interpret those facts – in order to reveal the alienation they express and, in so doing, to advance the emancipatory project of Left Hegelianism. How successful Marx was in implementing this programme is subject to debate. What is clear is that, as the focus of his theoretical work turned away from Hegelian philosophy towards political economy, history and politics, he became disabused of the idea that Adam Smith or any other classical economist had gotten political economy descriptively right. His life's project, thereafter, was to rework the conceptual apparatus of classical economics – more usually in its Ricardian, not Smithian, form – with a view to revealing the real 'laws of motion' of the capitalist mode of production. In this endeavour, Feuerbachian philosophical anthropology seemed to play no role. But, following the lead of Georg Lukacs (1923) several decades earlier, the Marxist humanists pointed out that there was, in *Capital*, an explicit point of connection – in the text on commodity fetishism. It was there that Marx brought his analysis of the commodity form to completion. But it was also there that, in modelling the commodity form, Marx identified the objectification of essential human traits in the process of capital accumulation. In consequence, capital, becomes a 'fetish', a *god* in Feuerbach's sense – one who controls economic behaviour by force of (illusory) power.

Structuralist Marxists, like Louis Althusser, were intent on reading Left Hegelianism out of the Marxist canon. They therefore treated Marx's references to fetishes and gods as ironic figures of speech, even as they attempted to enlist the text on commodity fetishism in the service of opposition to Marxist humanism. Borrowing a concept from the French philosopher Gaston Bachelard (1884–1962), Althusser (1965) disparaged Marx's early work by asserting the existence of an 'epistemological break' within the Marxist corpus. What he had in mind was roughly a 'paradigm shift' – not, however, within an ongoing scientific practice but between pre-scientific modes of thought and the inception of a new science. In Althusser's account, two

previously monumental epistemological breaks had occurred – one that established mathematics in ancient Greece, and one that established the sciences of nature in 17th-century Europe. Marx's achievement was supposedly on a par with these; he opened up a science of history. He did so by anticipating the structuralist turn the 'human sciences' (in France mainly) would later take – first in linguistics and psychology, later in anthropology and psychoanalysis. Specifically, in *Capital* and other writings of his maturity, Marx explained a range of diverse 'surface' phenomena by construing them as effects of the workings of a relatively small number of underlying, generally invariant 'deep' structures. The text on commodity fetishism lent itself to this construal of Marx's explanatory practice in as much as it depicted the perceptions of economic agents as effects of the unseen but causally efficacious process of surplus value extraction. Thus Marx's account can be seen as a theory of necessary (systemically induced) *misperception* – consonant with notions of explanation that contemporaneous structuralists endorsed. Perhaps the most innovative use Althusser made of commodity fetishism was in his theory of ideology, according to which modes of production constitute experiential subjectivity by 'interpellating' the human subjects who support or 'bear' them.

We now inhabit a different intellectual universe. In the past several decades, it has come to be widely believed, by erstwhile Marxists as much as by 'bourgeois economists', that Marx's focus on production rather than exchange inhibited the development of analytical economic tools. In so far as this belief is sound, the emphasis Marxists placed on commodity fetishism is partly to blame. The explanatory strategies of Marxist humanists and of structuralists have fallen into disrepute, too – largely because, in both cases, though for different reasons, the alleged connections between appearance and reality were never satisfactorily explained. No sustainable account was given either of how interpretation should proceed in the Marxist humanist case or, in the structuralist case, of how deep structures can be discerned in surface phenomena. Thus, commodity fetishism has fallen on hard times. However, we should not conclude that there is nothing viable in the concept or in the theoretical traditions that, until recently, magnified its importance. Hegelianism certainly, and structuralism possibly, still have much to teach us. The last word may not yet have been said on the theory of surplus value, either. If and when interest in Marx resumes, it will certainly be useful to revisit these issues. The notion of commodity fetishism played a key role in mid- and late 20th-century Marxism. The core idea it articulates – that necessary misperceptions sustain the capitalist order – can again provide useful insights. The concept may not be forever doomed to be of historical interest only.

See Also

- ▶ Capitalism
- ▶ Marx's Analysis of Capitalist Production

Bibliography

Althusser, L. 1965. *For Marx*, 1969. New York: Pantheon Press.

Althusser, L. 2002. Ideology and ideological state apparatus. In *Lenin and philosophy and other essays.* New York: Monthly Review Press.

Althusser, L., and E. Baliber. 1968. *Reading capital.* London: New Left Books, 1970.

Cohen, G. 1988. The labour theory of value and the concept of exploitation. In *History, labour and freedom: Themes from Marx.* Oxford: Clarendon Press.

Feuerbach, L. 1841. *The essence of christianity.* Amherst: Prometheus, 1989.

Fromm, E. 1975. *Marx's concept of man*. New York: Continuum.

Lukacs, G. 1923. *History and class consciousness.* London: Merlin Press, 1967.

Marx, K. 1844. Paris manuscripts (also called economic and philosophical manuscripts). In *Collected works*, ed. Karl Marx and Frederick Engels, vol. 3. New York: International Publishers, 1975.

Marx, K. 1845. The German ideology. In *Collected works*, ed. Karl Marx and Frederick Engels, vol. 5. New York: International Publishers, 1976.

Marx, K. 1867. *Capital,* vol. 1. Moscow: Progress Publishers, 1965.

Roemer, J. 1981. *Analytical foundations of marxist economic theory.* New York: Cambridge University Press.

Steedman, I. 1981. *Marx after sraffa*, rev. edn. New York: Shocken.

Commodity Money

François R. Velde and Warren E. Weber

Abstract

Commodity money is a medium of exchange that may be transformed into a commodity, useful in production or consumption. Although commodity money is a thing of the past, it was the predominant medium of exchange for more than two millennia. Operating under a commodity money standard limits the scope for monetary policy, actions that alter the value of money. However, it does not eliminate monetary policy entirely. The value of money can be altered by changing the commodity content or legal tender quality of monetary objects, or by restricting the conversion of commodities into money or vice versa.

Keywords

Banking; Bank of England; Bimetallism; Brassage; Bretton woods system; Coinage; Commodity money; Convertibility; Debasement; Discount rate; Fiat money; Foreign exchange markets; Gold standard; Great depression; Inflation; Monetary policy; Money changers; Seigniorage

JEL Classifications

E4

A commodity is an object that is intrinsically useful as an input to production or consumption. A medium of exchange is an object that is generally accepted as final payment during or after an exchange transaction, even though the agent accepting it (the seller) does not necessarily consume the object or any service flow from it. Money is the collection of objects that are used as media of exchange. Commodity money is a medium of exchange that may become (or be transformed into) a commodity, useful in production or consumption. This is in contrast to fiat money, which is intrinsically useless.

Commodity money can also be thought of as a medium of exchange that contains an option to consume a predetermined service flow at little or no cost. The option can be exercised in various ways, depending on the object. Coins can be melted down (at little cost) and the metal applied to non-monetary uses. In the case of paper or token money under a commodity money standard, the medium of exchange itself is intrinsically useless, but it is costlessly convertible into a specified quantity of the commodity on demand. Fiat money can also be converted into goods or services, but in quantities that will depend on market prices.

Commodity money is a thing of the past; countries worldwide now use fiat money standards. However, this is a relatively recent development. Commodity money, primarily in the form of coined metals, was the predominant medium of exchange for over two millennia. Although operating under a commodity money standard limits the scope for monetary policy, it does not eliminate it entirely. The history of commodity money is replete with numerous ways in which governments have altered the monetary system to achieve various goals.

From Commodity Money to Fiat Money

In early or primitive societies, it is often difficult to characterize the general patterns of trades and transactions, let alone determine how generally accepted a particular commodity might be. Nevertheless, a wide range of commodities have been reportedly used as money (cowry shells, wampum, salt, furs, cocoa beans, cigarettes and so on), perhaps the most exotic being the stone money of the island of Yap in Micronesia.

General acceptability of monetary objects is most clearly ascertained when the objects are standardized and exchanged repeatedly. With metallic commodities, the standardized objects are called coins. Coinage of metal began in the eastern Mediterranean region or the Middle East, India and China between the 6th and 4th centuries

BC. Coinage has developed in parallel and broadly similar ways in these areas.

The metals most commonly used have been gold, silver and copper (in decreasing order of scarcity), in varying degrees of fineness (silver mixed with substantial amounts of copper, called billon). Lead, tin and various copper alloys (bronze, brass, potin) have also been used, although less frequently than the more common metals. The metal is either mined or acquired through trade. The most common method of coinage is striking with a die, although cast coins are also found. In many legal traditions the right of coinage is a prerogative of the public or central authority, although it may be delegated or leased to regional authorities or private parties. This prerogative may also extend to mining. In other words, the rules governing the supply of commodity money vary from government monopoly to minimal regulation.

In Europe and the Mediterranean, coinage – an invention mythically linked to Croesus, King of Lydia – began near the Aegean Sea in the 6th century BC. The use of money developed considerably in Greek and Roman times, leading to a three-tiered system of gold, silver, and copper denominations. In the Roman empire, the provision of coinage was a government monopoly. The collapse of the empire in the West led, after a long transition, to a purely silver-based monetary system, with a largely decentralized provision of minting. Uniformity of coinage was restored under Charlemagne but quickly disappeared along with political fragmentation. Gold returned in common use from the mid-13th century. By the 14th century, most mints in western Europe operated along similar lines, with more or less unrestricted coinage on demand provided by profit-making mints. A great multiplicity of monetary systems persisted, giving rise to both foreign exchange markets (the earliest financial markets) and money changers (the first financial intermediaries).

The first instances of token coinage (coins that are intrinsically useless but are claims to fixed amounts of the commodity) appeared in the 15th century in Catalonia. Notes convertible on demand appeared in the 17th century, in Sweden and later in England. For a more complete discussion of medieval European coinage, see Spufford (1988).

Coins appear to have been used in India in the early 4th century BC and were probably used before then. The earliest coins were so-called punch-marked coins and were adaptations of Greek prototypes. Coins were first used in China and the Far East about the same time as in India. The distinctive bronze coinage with the square hole in the middle first appeared in the 3rd century BC. Early coins in eastern Islamic lands were copies of Byzantine gold and bronze coins; those in the East were copies of Sassanian silver coins. For more on coinage in India and the Far East, see Williams (1997).

Until the 19th century, coins typically bore no indication of face value, and their market value could fluctuate even relative to one another. From the late Middle Ages, governments increasingly sought to regulate the value of coins in some manner, in particular assigning face value or legal tender value by decree. It became desirable to turn the collection of objects used as a medium of exchange into a stable system with fixed exchange rates between the objects. This was achieved to a large degree with bimetallism, a system in which gold and silver coins remained concurrently in circulation at a constant relative price. Its heyday was the mid-19th century, but beginning in 1873 the system was quickly abandoned, and by the First World War countries were using either gold only or (in Africa and eastern Asia) silver only. (Bimetallism is discussed in more detail in Redish 2000, and Velde and Weber 2000.) The development of banking in the 19th century also led to increased use of (convertible) notes and other monetary instruments.

The First World War brought about the suspension of convertibility of the notes in many countries. Most countries returned to convertibility between 1926 and 1931, but the onset of the Great Depression reversed the movement. After the Second World War the only major country whose currency was in any way directly tied to a commodity was the United States under the Bretton Woods system: dollars were convertible

by non-residents of the United States into gold on demand, while other currencies of the system were convertible into dollars. The link between gold and the dollar was severed in 1971. Fiat money standards are now universal.

The Nature of Commodity Money

The definitions of commodity and fiat monies given above make it seem as if there is a clear distinction between the two. It is more helpful, however, to think of media of exchange along a continuum. An object serving a purpose as a medium of exchange has value above its intrinsic content, reflecting the value of the service as a medium of exchange.

Because the value of a commodity qua commodity and the value as a medium of exchange can differ, the value of all commodity monies has a fiat component. A pure fiat money is one for which this fiat component makes up its entire value. A nice theoretical discussion of commodity and fiat monies is given by Sargent and Wallace (1983).

Price-Level Determination

It is natural that the medium of exchange in an economy is what becomes the unit of account, the unit in which debt contracts and the prices of goods and services are expressed. It is natural because the money appears on one side of virtually every transaction.

Because commodity money has an intrinsic value apart from that which it obtains by being a medium of exchange, its relative price will not be zero. Thus, in a commodity money economy, the value of money (the inverse of the price level) is bounded away from zero. Moreover, in a canonical commodity money system (see below) with unlimited minting at a set price, the value of money and its quantity tend to remain within a band. If the value of money falls far enough, it becomes preferable to exercise the option and convert some of it into other, non-monetary uses, thus reducing the quantity and preventing the value from falling further. Conversely, if the value of money rises high enough, it becomes worthwhile for agents to turn metal into coins at the mint at the set price, thus increasing the quantity of money. Such a self-regulating commodity money system provides an anchor to the price level. This has been touted as one of the advantages of a commodity money system, particularly in the case of the gold standard.

The question of price-level determination becomes more complicated when multiple commodity monies are made out of different commodities. An example is the circulation of full-bodied gold and silver coins. Should the unit of account be the gold coin or the silver coin? This matters because under a commodity money system a monetary authority does not have the ability to set the exchange rate between monies of different commodities forever. Thus, to the extent that the unit of account is used in contracts to determine the amount of future payments, the choice of the unit of account can affect the allocation of goods and services. This was one of the issues surrounding the possible adoption of a bimetallic standard mentioned above.

The inability of the monetary authority to set the exchange rate between different monies goes away under a pure fiat money system. Because fiat money is (virtually) costless to produce, the monetary authority can costlessly exchange one money for another to maintain whatever exchange rate is desired between different monies that it issues.

Monetary Policy

The fact that a commodity is used as money alters its value. This is because part of the total quantity of the commodity – namely, the metal locked up in the form of coins, or the reserves held by the monetary authority – is not available for non-monetary uses. The allocation between monetary and non-monetary uses is determined in equilibrium. Restrictions on the ability to change this allocation, such as restrictions on melting or exporting coins, or limitations on the minting of metal, will have an effect on the equilibrium value of the money even if it has no immediate effect on

the allocation itself. (Since money is an asset, its valuation is forward looking.) Thus, there is scope for monetary policy under a commodity money standard, although what constitutes monetary policy is different from and more limited in scope than what holds under a pure fiat money standard.

Monetary policy consists in actions that tend to alter the value of money. In a commodity money system, the value of money is the value of the option we have described. (The strike price of the option is zero, since the commodity is the money.) Most aspects of monetary policy with commodity money consist in modifying this option, typically by modifying the institutions governing the exercise of the option rather than by modifying the quantity of money, which the authority usually cannot control directly. When the monetary authority is directly involved in the provision of the money, it may directly profit from its actions. Potential profit is often an important consideration of monetary policy.

The canonical form of a commodity money standard comprises the following. One or more commodities are chosen to be the standard to which the monetary system will be anchored. The monetary authority defines the specifications of the monetary objects (weight, fineness) and defines the unit of account in terms of these monetary objects. The conversion of commodity into commodity money and vice versa is as costless as possible. In particular, the monetary authority provides for unlimited (and even costless) conversion of the commodity into monetary objects (coins or notes). Conversely, it places no hindrances on the conversion of monetary objects into commodities (coins can be melted, notes are convertible on demand), nor does it place limitations on the consumption of the commodity or its service flow (free possession, unrestricted import and export of the commodity). The monetary objects are unlimited legal tender.

One type of monetary policy modifies the specifications of monetary objects and units of account. An example is debasement, which is reducing the commodity content of a monetary object (and, frequently, of the corresponding unit of account). The result of debasement is inflation, since nominal prices will be adjusted to maintain the relative prices of goods and money. And, just as occurs with fiat money, inflation has the effect of transferring wealth from nominal creditors to nominal debtors. Since governments generally tended to be debtors, debasements were used to reduce the amount of their debts. Historically, debasements also had the secondary effect of increasing seigniorage revenue, since the quantity of coins minted tended to increase significantly after debasements that involved the introduction of new coins (see Rolnick et al. 1996; Sargent and Smith 1997). Debasements were also used by governments to remedy malfunctions of a multiple-denomination commodity money system (see Sargent and Velde 2002).

A second type of monetary policy adds or modifies restrictions on the conversion of commodity into money or money into commodity. For example, minting might be restricted by quantity, in which case the authority decides how much to mint. Minting might be unlimited but subject to a fee, called seigniorage. Governments typically charged such a fee, both to cover the actual costs of minting (called brassage) and as a tax (England was the first, in 1666, to provide minting at no cost). The rate of this tax or, equivalently, the price paid by the mint for bullion might be changed. These restrictions tended to alter the allocation of the commodity between monetary and nonmonetary uses, and hence the value of the commodity and the money.

A third type of monetary policy sets limits to the legal tender quality of certain coins, or changes their legal tender value. Since coins did not have face values until the 19th century, it was up to monetary authorities to set, and from time to time alter, the legal tender values of coins. Frequently, foreign coins were authorized as legal tender at rates set for domestic coins. Countries attempting to maintain bimetallism in the face of fluctuations in the relative price of gold and silver often had to adjust the face value of either their gold or silver coins. Changes in the legal tender values could also be motivated by fiscal considerations or by attempts to target a particular price level or exchange rate.

The physical nature of the medium of exchange led to a particular set of concerns.

Coins, like anything else, depreciate with use, through wear and tear. Since coins of different values have different usage rates, the depreciation rate varied by denomination. Also, being roughly constant over time, depreciation depended on the age of the coin. Finally, imperfect minting technology as well as actions by the public (clipping, sweating) aggravated the disparities between coins. This factor introduced heterogeneity among coins and hindered the achievement of a stable and uniform monetary system. Improvements in coin production partially remedied the problem, as did periodic recoinages.

When the monetary objects consist not only of coins but also of paper currency or tokens that are demand promises to the commodity, a fourth type of monetary policy is available: suspension of convertibility. The monetary authority can refuse to honour the promise of convertibility for some period of time. An example is the suspension of convertibility by the Bank of England between 1797 and 1819 during the wars with France. During the 19th century suspensions were not uncommon during financial or fiscal emergencies, with the understanding that the suspension would end after the emergency and convertibility would be restored at the preexisting parity. This understanding has been described as a state-contingent gold standard (see Bordo and Kydland 1996).

When there is a central bank, an additional monetary tool is to change the discount rate, the interest rate at which the central bank lends reserves to the banking system. During the gold standard period, this was the primary means by which central banks affected the exchange rate of their money against the monies of other countries.

Conclusion

Commodity money is a thing of the past; countries worldwide now use fiat money standards. This practice has led to an efficiency gain in the sense that resources that were once tied up in coins are now available for consumption and production (perhaps prompting John Maynard Keynes to refer to gold as the 'barbarous relic').

It has also led to a greater scope for monetary policy because the supply of money can be changed almost costlessly. However, along with this greater scope has come the greater potential for governments to use inflation to collect seigniorage revenue or to reduce the real value of their debts. How to use the freedom that commodity money restricted is still a matter of debate.

See Also

- ▶ Bimetallism
- ▶ Bretton Woods System
- ▶ Fiat Money
- ▶ Gold Standard

Bibliography

Bordo, M., and F. Kydland. 1996. The gold standard as a commitment mechanism. In *Modern perspectives on the gold standard*, ed. T. Bayoumi, B. Eichengreen, and M. Taylor. Cambridge: Cambridge University Press.

Kiyotaki, N., and R. Wright. 1989. On money as a medium of exchange. *Journal of Political Economy* 97: 927–954.

Luschin von Ebengreuth, A. 1926. *Allgemeine Münzkunde und Geldgeschichte des Mittelalters und der neueren Zeit*. Munich: R. Oldenbourg.

Redish, A. 2000. *Bimetallism: An economic and historical analysis*. Cambridge: Cambridge University Press.

Rolnick, A., F. Velde, and W. Weber. 1996. The debasement puzzle: An essay on medieval monetary history. *Journal of Economic History* 56: 789–808.

Sargent, T., and B. Smith. 1997. Coinage, debasements, and Gresham's laws. *Economic Theory* 10: 197–226.

Sargent, T., and F. Velde. 2002. *The big problem of small change*. Princeton: Princeton University Press.

Sargent, T., and N. Wallace. 1983. A model of commodity money. *Journal of Monetary Economics* 12: 163–187.

Spufford, P. 1988. *Money and its use in medieval Europe*. Cambridge: Cambridge University Press.

Sussman, N., and J. Zeira. 2003. Commodity money inflation: Theory and evidence from France in 1350–1436. *Journal of Monetary Economics* 50: 1769–1793.

Velde, F., and W. Weber. 2000. A model of bimetallism. *Journal of Political Economy* 108: 1210–1234.

Williams, J., ed. 1997. *Money: A history*. New York: St Martin's Press.

Commodity Reserve Currency

Albert Gailord Hart

Commodity Reserve Currency (CRC for short) is a proposal for re-establishing an international monetary 'standard' – basing it upon a 'basket' of widely used commodities. Recent experience shows the inconvenience of lacking a standard. While restoration of a gold standard has many supporters, gold has become so remote from the goods-and-services economy that for decades now governments have had scope to play tricks with its price; and since the early 1970s, the value of gold has been highly unstable. Can there be a commodity standard other than gold, less abstract and linked to articles of everyday use? A good way to study this question is to examine the feasibility and desirability of CRC.

Pioneers of the CRC proposal were Jan Goudriaan and Benjamin Graham. Variant proposals have come from a number of economists, including Lord Keynes and Friedrich Hayek. The nearest approach to a standard version is probably still the submission of 1964 to the United Nations Conference of Trade and Development by Albert Hart, Nicholas Kaldor and Jan Tinbergen. No governments or major multilateral bodies have sponsored CRC, though there has been official and private use of 'baskets' of currency units (ECU, etc.), and of 'baskets' of securities (traded on various private commodity exchanges).

Advocates of CRC propose that it be administered by a multinational agency, which we may call IMA – presumably to be a branch of the International Monetary Fund. A currency unit (which following Lord Kaldor we may call the Bancor) is to be defined as the value of a basket of primary commodities, with fixed physical composition.

Like the administrator of a traditional gold standard, IMA must buy or sell at a stated price (plus or minus a margin to avoid a hair-trigger effect) as much of the monetary commodity (i.e. the commodity baskets) as may be offered or demanded. As to the margin, some CRC proponents suggest 5 per cent on each side of par; but a wider range would have great advantages. Prices of the individual component commodities could fluctuate more than the basket. Any one commodity price could rise; but if its rise would bring the basket above the posted selling price, sales of baskets by IMA would bring a compensating fall of other prices.

The price of a basket with fixed physical composition is an index number of commodity prices, weighted by the quanta of the various items included. Hence CRC may be viewed as a scheme to stabilize an index of primary-commodity prices. CRC thus would have a counter-cyclical effect – holding within bounds the fluctuation of income for the world's exporters and producers of primary commodities, and by the same token the fluctuation of major elements in the world's cost of living.

An effective CRC would require that national currencies be tied to the Bancor by fixed exchange rates, or at least by not-too-movable pegs. The general stabilization effect of the proposal would vanish if major currencies were allowed to float against the Bancor.

Many primary commodities have been proposed for a CRC basket. Major criteria for inclusion are:

(1) *Standardization*, of the sort necessary to run futures markets on a commodity exchange – with rules for dealing with quality differences.
(2) *Storability* for at least a year or two without excessive cost or loss of quality. Security against fire, looting, requisition by local governments, etc., is implied.
(3) *Improbability of major price manipulations* by governments or by combinations of producing enterprises. Long before OPEC rose to power in 1973, this criterion led advocates of CRC to omit petroleum from the proposed basket.
(4) *Reliability of commodity contracts*, enabling the IMA to replace physical holdings with contracts for future or spot delivery if this will reduce costs. Use of futures could be of

humanitarian importance in case of shortages of foodstuffs such as rice.

These criteria would admit most of the world's important grains, fibres, fats and oils, beverage crops, primary metals – and probably a number of basic chemicals and forest products. Amounts of the various commodities in the basket would reflect their weight in world production and/or trade.

The CRC basket must be large (worth several million US dollars), because for efficient trading each element must be a multiple of a wholesale lot. This large basket-size would be appropriate, since CRC is designed as a vehicle for holding national monetary reserves rather than for retail or even wholesale trade.

IMA holdings of the various commodities must be parcelled out to points of delivery and storage along the normal trade routes of the commodities. Correspondingly, IMA purchases and sales must be handled by agents at the various trading points. To tell whether at any moment baskets must be bought or sold, IMA must sum up bids or offers reported by these agents. IMA must take the initiative – instructing all agents to sell or buy – whenever the sum of bid prices for the elements of the basket adds up to the posted selling price for the basket, or the sum of asked prices to the buying price.

Objections to CRC have hinged primarily on costs and/or on the difficulties of getting a CRC system under way. There has been continuing debate about the size of the reserve needed to validate IMA selling offers, about the cost of holding stocks of various suggested commodities, etc. Such costs must be compared with benefits from reducing cyclical fluctuations in primary producers' incomes, etc. – and of doing so without engaging in commodity-by-commodity operations. Benjamin Graham used to stress that the success of such operations hinges on restriction of commodity production, whereas success of CRC would stimulate production. For this and other reasons, the cost/benefit problem is complex.

During episodes of worldwide commodity stringency, accumulation of a commodity reserve has seemed impossible. When there has been a great piling up of stocks (as at this writing in early 1986), it has seemed as if mobilization of stocks held by the European Economic Community, the United States, various other governments such as Brazil, cartels such as that for tin, and private concerns such as copper companies, might permit a very rapid start.

Starting a CRC during a period of widespread shortages could be highly inflationary; starting it during a general economic downswing could mitigate a recession. There is debate as to whether accumulation of a reserve must stick to previously agreed proportions of the different commodities, or whether (as proposed by Keynes and more recently by Kaldor) the IMA should buy individual commodities at such dates and prices as seem wise. On this route, the composition of the CRC basket and the selling-price offer would grow out of the process of accumulation. This variant, plainly, would increase the similarity of IMA's operations to those of commodity 'stabilization' groups and remove the impersonality held to be the central virtue of the Goudriaan/Graham scheme.

See Also

- ▶ Currencies
- ▶ Gold Standard
- ▶ International Monetary Policy

References

Bennett, M.K., et al. 1949. *International commodity stockpiling as an economic stabilizer*. Stanford: Stanford University Press.

Goudriaan, J. 1932. *How to stop deflation*. London: The Search Publishing Company.

Grubel, H.G. 1965. The case against international commodity reserve currency. *Oxford Economic Papers* 17(1): 130–135.

Hallwood, P. 1986. External economy arguments for commodity stockpiling: A review. *Bulletin of Economic Research* 38(1): 25–41.

Harmon, E. 1959. *Commodity reserve currency*. New York: Columbia University Press.

Hart, A.G., N. Kaldor, and J. Tinbergen. 1964. *The case for an international Commodity Reserve Currency*.

Geneva: United Nations Conference on Trade and Development. 17 February 1964. E/CONF.46/p/7. Conveniently accessible in N. Kaldor, *Essays on economic policy–II.* London: Gerald Duckworth & Co., 1964.

Hayek, F.A. 1984. The future monetary unit of value. In *Money in crisis*, ed. N. Barry. Siegel: Pacific Institute for Policy Research.

Newbery, D.M.G., and J.E. Stiglitz. 1981. *The theory of commodity price stabilization*. Oxford: Clarendon.

Common Factors

Heather M. Anderson

Abstract

This article outlines and illustrates several types of common factor models that are found in the applied economics literature. These factor models include those based on principal components, classical factor analysis, dynamic factor analysis and common features, and the discussion addresses the identification and estimation of factors, as well as the use of common factor models.

Keywords

ARMA processes; Autoregressive moving average (ARMA) processes; Canonical correlations; Capital asset pricing model; Classical factor models; Coincident indices; Common factors; Common feature models; Common trend model; Diffusion indexes; Dynamic factor (or index) models; Factor analysis; Kalman filter; Principal component analysis; Real business cycle models; Reduced rank regressions; Static factor models; Stone, J; Term structure of interest rates; Time series analysis

JEL Classifications

C32

Economic analysis frequently involves the study of variables that exhibit similar behaviour, and it is often of interest to model this comovement. Well-known examples of comovement in multivariate data sets include business cycles in macroeconomic indicators and shifts in the entire term structure of interest rates, and researchers sometimes attribute this comovement to a small set of underlying forces or latent 'factors' that influence each variable in the system. It is then convenient to think of the variation in each variable in the system as the sum of two types of (unobserved) components, one of which captures variation that is due to 'common factors', while the other captures all other variation. Models that attribute comovement to common factors are called common factor models, and common factor analysis involves the identification and study of the common factors.

Common factor models are particularly popular in empirical settings because they offer parsimony, and simplify estimation by reducing the number of parameters that need to be estimated. Economists will typically be interested in interpreting common factors so that they can explain why comovement occurs. Economic theory sometimes predicts common factors. Perhaps the best-known example of this is the capital asset pricing model, in which the (excess) return for the market portfolio is the common factor in the (excess) return for each individual stock. Another well-known example arises when the term structure of interest rates is modelled, because the no arbitrage condition implies that the entire term structure is determined by a single factor, which is the instantaneous interest rate.

A simple model that captures the concept of common factors in a set of N time-series in the (demeaned) vector $Y_t = (Y_{1t}, Y_{2t}, \ldots, Y_{Nt})'$ is given by

$$Y_t = AF_t + \varepsilon_t, \tag{1}$$

where F_t is an $r \times 1$ vector that contains r common factors, A is an $N \times r$ factor loading matrix (with $rank(A) = r < N$), and ε_t contains N idiosyncratic components. With the use of Σ_Y, Σ_F and Σ_ε to denote the variance covariance matrices of Y_t, F_t and ε_t, it is usual to assume that Σ_ε is diagonal, and it is also common to normalize the set of r factors in F_t by assuming that $\Sigma_F = I_r$.

Model (1) is similar to conventional factor models that are often used in cross-sectional settings, although the variables are specified here as time series, to facilitate discussion on dynamic factor models. If there is no serial correlation in Y_t or F_t, or if estimation is undertaken as if this is the case, then (1) is called a static factor model. It is usual to assume that F_t and ε_t are jointly stationary, that $E(\varepsilon_t) = 0, E(F_t\varepsilon_t') = 0$, and that ε_t contains no serial dependence, but these latter assumptions can be relaxed, depending on the type of factor model under consideration.

There are many ways to identify the factors in (1), and standard techniques include the use of principal component analysis, factor analysis and canonical correlations to estimate the parameters of various associated reduced rank regressions. More recently, researchers have focused on the time series properties of multivariate data-sets, and modern factor models include dynamic factor (or index) models, and models that incorporate common features. These latter models incorporate various ways of allowing the factors to follow specific dynamic processes, or to contain specific time-series properties.

Principal Component Models

Principal component analysis involves the intuition that most of the variance in Y_t will be attributable to variance in the r components in F_t. The factors F_t are modelled as linear combinations of Y_t, and their identification is based on finding the r (orthogonal) linear combinations of Y_t that have the most variance. In practice this involves finding the eigen values $\lambda_1 > \lambda_2 > \ldots > \lambda_N$ and associated eigenvectors $f_1, \ldots, f_N$ of the form $f_i = \beta_i' Y_t$ that are associated with the roots of the equation $|\widehat{\Sigma}_Y - \lambda I| = 0$, where $\widehat{\Sigma}_Y$ is an estimate of Σ_Y. The β_i are picked so that $\left(\widehat{\Sigma}_Y - \lambda_i I\right)\beta_i = 0$ and $\beta_i'\beta_i = 1$, and the factors F_t are then defined by $F_t = (f_1, \ldots, f_r)$. This decomposition ensures that $E\left(F_t'\varepsilon_t'\right) = 0$, but it implies that the ε_t (which are each linear combinations of $(f_{r+1}, \ldots, f_N)$) will be correlated with each other so that Σ_ε will not be diagonal. Principal components estimators of common factors and factor loadings are also the least squares estimators of the reduced rank regression given by

$$Y_t = \underset{(N\times r)}{A}\ \underset{(r\times N)}{B}\ Y_t + \varepsilon_t, \tag{2}$$

where BY_t contains the r factors $F_t = (f_1, \ldots, f_r)$. Anderson (1984, ch. 11) provides a standard reference.

In practice, one needs to determine r before estimating common factor models, and this is often based on the ratio given by $\frac{\lambda_{r+1}^2+\ldots+\lambda_N^2}{\lambda_1^2+\ldots+\lambda_N^2}$. This ratio measures the loss of information in the reduced rank system relative to an unrestricted system, and typically investigators will choose r so that this ratio is kept small. Bai and Ng (2002) have developed model selection criteria that are consistent as $\{N, T\} \to \infty$.

Principal components are usually used for dimension reduction, and economic interpretation of the resulting factors is rarely straightforward. However, Stone (1947) has summarized a set of series from the US national accounts, associating the first three principal components with income, income growth and time and Chamberlain and Rothschild (1983) have promoted the use of principal components for estimating approximate factor models of asset prices. Stock and Watson (2002) have suggested the use of diffusion indexes (principal component factors associated with large macroeconomic data sets) for forecasting key macroeconomic variables, and the interest here centres on using information in the factors rather than interpreting the factors themselves.

Classical Factor Models

Classical factor models are closely related to principal components models, but the underlying intuition and assumptions are different. In this case the key assumption is that Σ_ε is diagonal so that the ε_t describe idiosyncratic effects that are unique to each variable in Y_t, while the factors describe joint effects in Y_t. The assumptions that $E(\varepsilon_t) = 0$ and $E\left(F_t\varepsilon_t'\right) = 0$ still hold, and the ε_t are assumed to contain no serial dependence. Under these

assumptions $\Sigma_Y = A\Sigma_F A' + \Sigma_\varepsilon$, and estimates for A, Σ_F and Σ_ε can be found by maximizing the function

$$L_T(A, \Sigma_\varepsilon) = -\frac{T}{2}\ln|\Sigma_Y| - \frac{1}{2}\sum_{t=1}^{t=T} Y_t'\Sigma_Y^{-1}Y_t$$

subject to the condition that $rank(A) = r$ and a set of normalization restrictions that will uniquely identify the $(r+1)(N+\frac{1}{2}r)$ parameters. Researchers often use the joint restrictions that $\Sigma_F = I_r$ and that $A'\Sigma_\varepsilon^{-1}A$ is diagonal for normalization, but other normalizations are common (see Anderson 1984, ch. 14, for details). If Y_t and ε_t are normally distributed then L_T (A, Σ_ε) is the log likelihood for Y_t (if we ignore the constant term), but, even when Y_t and ε_t are not normally distributed, the maximization of $L_T(A, \Sigma_\varepsilon)$ delivers quasi-maximum likelihood estimates. There are several ways of using the estimates of A and Σ_ε to obtain estimates of the factors in F_t, and perhaps the best-known of these is Bartlett's (1937, 1938) method based on generalized least squares given by

$$\widehat{F}_t = \left(\widehat{A}'\widehat{\Sigma}_\varepsilon^{-1}\widehat{A}\right)^{-1}\widehat{A}'\widehat{\Sigma}_\varepsilon^{-1}Y_t.$$

As above, it is necessary to determine r prior to estimating the factors, and, on the assumption of normality, the likelihood ratio test statistic for testing H_0: $r = s$ versus H_A: $r > s$ is given by

$$-T\left[\ln|\widehat{\Sigma}_Y\right| - \ln|\widehat{A}\widehat{A}' + \widehat{\Sigma}_\varepsilon|] = -T \quad \Sigma_{i=s+1}^{i=N}\ln\left(1-\widehat{\lambda}_i^2\right),$$

where the $\widehat{\lambda}_i$ are the characteristic roots of $\widehat{A}'\widehat{\Sigma}_\varepsilon^{-1}\widehat{A}$ (in decreasing order) and $\hat{A}$ is estimated under the null. The test statistic is asymptotically distributed as a χ_q^2 with $q = [(N-s)^2 - N - s]/2^\circ$ of freedom under the null.

There are numerous applications of classical factor analysis to economic problems, and an early example includes Stone's (1945) factor analysis of the demand for N commodities. Another example includes a factor model of returns by Deistler and Hamann (2005).

Dynamic Factor Models

Classical factor models are not well suited to multivariate analysis of time series because they assume no serial correlation in ε_t, and, if there are any dynamics in F_t, then they are implicit and not explicitly modelled. Dynamic factor models address these concerns by treating the ε_t and F_t as autoregressive moving average (ARMA) processes. The innovations that underlie the N processes for ε_t are assumed to be mutually uncorrelated, and uncorrelated with the innovations that underlie the F_t at all leads and lags, but the factors themselves can be mutually correlated. Different variables in Y_t can then move together because they are functions of the same factor (s), or because they are functions of different factors that are themselves correlated.

The identification and estimation of small-scale dynamic factor models is sometimes based on spectral techniques (see Geweke 1977; or Sargent and Sims 1977), and use of the Kalman filter in the time domain (as in Engle and Watson 1981, or Harvey and Koopman 1997) provides an alternative approach. Dynamic factor models have been particularly popular for estimating factor models of business cycles (as in Geweke and Singleton 1981), but they have also been used for studying the term structure (Singleton 1980) and fluctuations in employment across different industrial sectors (Quah and Sargent 1993).

Recent work has shifted towards the identification and estimation of common factors in large-scale models, relying on the use of large N to obtain consistent estimates of the factors. One strand of this literature adopts a static framework and standard principal components to estimate the factors, and then builds dynamic models of the factors. The resulting models are sometimes called approximate dynamic factor models. Applications of this approach include Stock and Watson's diffusion index (2002), and Bernanke and Boivin's (2003) estimation of a monetary policy reaction function. Another strand of this literature allows different variables to depend on different lags of common factors. These 'generalized dynamic factor models' are estimated using 'dynamic principal components', which are the principal components of

spectral density matrices at different frequencies. Applications of this latter approach include a study of business cycle dynamics in the United States (Forni and Reichlin 1998) and the development of a coincident index for Europe (Forni et al. 2000).

Canonical correlation-based models

Principal component and factor models assume that the factors are linear combinations of the N variables in Y_t, but sometimes it is useful to assume that the factors are linear combinations of M variables contained in another multivariate time series denoted by X_t. The variables in X_t will often include lags of the variables in Y_t, but X_t can also include variables that would be classified as explanatory variables in a regression context. The factors in (1) can now be written in the form $F_t = B'X_t$ (where $rank(B) = r < \min\{N, M\}$). In what follows, we assume that the ε_t in (1) are white noise and uncorrelated with X_t.

The main idea behind a canonical correlations approach is to find linear combinations of X_t that are strongly correlated with linear combinations of Y_t, and, as for principal component models, the estimators of common factors and factor loadings are the least squares estimates of a reduced rank regression. In this case the regression is

$$Y_t = \underset{(N\times r)}{A}\ \underset{(r\times M)}{B}\ X_t + \varepsilon_t, \tag{3}$$

and the factors and factor loadings are related to the r largest roots of $R = \Sigma_Y^{-\frac{1}{2}}\Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY}\Sigma_Y^{-\frac{1}{2}}$, which is the multivariate generalization of the (squared) correlation coefficient between two variables. If we order these roots (also called squared canonical correlations) so that $\lambda_1^2 > \lambda_2^2 > \ldots > \lambda_r^2$, and let $V_1, V_2, \ldots V_r$ be the r associated eigenvectors, then the factor loadings and factors are given by $A_i = \Sigma_Y^{-\frac{1}{2}}V_i$ and $B_iX_t = \Sigma_Y^{-\frac{1}{2}}\Sigma_{YX}\Sigma_X^{-1}V_iX_t$. Anderson (1984, ch. 12) provides a detailed discussion of canonical correlations, while Izenman (1980) discusses the associated reduced rank regressions. When the variables in X_t are simply lags of the variables in Y_t, then the first factor is the best predictor of Y_t based on past history, the second factor is the next best predictor, and so on, and the factors provide a set of leading indicators for Y_t. When X_t consists of explanatory variables for Y_t, then the factors are often called coincident indices. One can base a test of $H_0 : r = s$ versus $H_A : r > s$ on the test statistic $-T\Sigma_{i=s+1}^{i=N}\ln\left(1-\widehat{\lambda}_i^2\right)$, which has a χ^2 distribution with $(m - s)(n - s)$ degrees of freedom under the null.

Common Feature Models

Common feature models are a special class of factor models in which the common factors have a statistical characteristic of interest, while the idiosyncratic components fail to have this characteristic. Common features were first introduced by Engle and Kozicki (1993) when they discussed serial correlation features – a situation in which each of N variables is serially correlated, but there are linear combinations that are white noise. Here, the presence of $N - r$ white noise linear combinations implies a factor model in which there are r serially correlated factors (which are sometimes called common cycles). An earlier example of a common feature model is Stock and Watson's (1988) common trend model which is valid when variables are cointegrated (as in Engle and Granger 1987). In this case the common factors are integrated of order one but the remaining components (often called error correction terms) are stationary. Other examples of common features include Vahid and Engle's (1993) common trend–common cycle representation, and common nonlinearity (Anderson and Vahid 1998).

The identification of common features involves finding linear combinations of the data that do not have the feature, and this can be done using a canonical correlations approach in which the variables in X_t model the characteristic of interest. To illustrate, lags of Y_t are put into X_t when testing for serial correlation features in Y_t, and lagged levels are included in X_t when testing for common trends. Factors associated with the lowest eigen values define linear combinations

that do not contain the feature, while factors associated with the highest eigen values are used to model the common features. Johansen's (1988) procedure provides a well-known example of this, although inference in this case is based on non-standard (rather than χ^2) distributions because the factors are non-stationary.

A well-known example of a common feature model is the real business cycle model of King et al. (1988), in which a common factor (productivity) generates the trend in output consumption and investment, and another factor (the deviation of capital stock from steady state) generates the common cycle.

See Also

- ▶ Reduced Rank Regression
- ▶ Time Series Analysis

Bibliography

Anderson, T. 1984. *An introduction to multivariate statistical analysis*, 2nd ed. New York: Wiley.

Anderson, H., and F. Vahid. 1998. Testing multiple equation systems for common nonlinear components. *Journal of Econometrics* 84: 1–36.

Bai, J., and S. Ng. 2002. Determining the number of factors in approximate factor models. *Econometrica* 70: 191–221.

Bartlett, M. 1937. The statistical conception of mental factors. *British Journal of Psychology* 28: 97–104.

Bartlett, M. 1938. Methods of estimating mental factors. *Nature* 141: 609–610.

Bernanke, B., and J. Boivin. 2003. Monetary policy in a data rich environment. *Journal of Monetary Economics* 50: 525–546.

Chamberlain, G., and M. Rothschild. 1983. Arbitrage, factor structure and meanvariance analysis in large asset markets. *Econometrica* 51: 1305–1324.

Deistler, M., and E. Hamann. 2005. Identification of factor models for forecasting returns. *Journal of Financial Econometrics* 3: 256–281.

Engle, R., and C. Granger. 1987. Cointegration and error correction representation, estimation and testing. *Econometrica* 55: 251–276.

Engle, R., and S. Kozicki. 1993. Testing for common features (with discussions). *Journal of Business and Economic Statistics* 11: 369–395.

Engle, R., and M. Watson. 1981. A one-factor multivariate time series model of metropolitan wage rates. *Journal of the American Statistical Association* 76: 774–781.

Forni, M., and L. Reichlin. 1998. Let's get real: A factor-analytic approach to disaggregated business cycle dynamics. *Review of Economic Studies* 65: 453–473.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin. 2000. The generalized dynamic factor model: Identification and estimation. *Review of Economics and Statistics* 82: 540–554.

Geweke, J. 1977. The dynamic factor analysis of economic times-series models. In *Latent variables in socioeconomic models*, ed. D. Aigner and A. Goldberger. Amsterdam: North-Holland.

Geweke, J., and K. Singleton. 1981. Maximum likelihood 'confirmatory' factor analysis of economic time series. *International Economic Review* 22: 133–137.

Harvey, A., and S. Koopman. 1997. Multivariate structural time series models. In *Systematic dynamics in econometric and financial models*, ed. C. Heij, H. Schumacher, and C. Praagman. Chichester: Wiley.

Izenman, A. 1980. Assessing dimensionality in multivariate regression. In *Handbook of statistics*, vol. 1, ed. P. Krishnaiah. Amsterdam: North-Holland.

Johansen, S. 1988. Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control* 12: 231–254.

King, R., C. Plosser, and S. Rebelo. 1988. Production, growth and business cycles II. New directions. *Journal of Monetary Economics* 21: 309–341.

Quah, D., and T. Sargent. 1993. A dynamic index model for large cross sections. In *Business cycles, indicators and forecasting*, ed. J. Stock and M. Watson. Chicago: University of Chicago Press.

Sargent, T., and C. Sims. 1977. Business cycle modeling without pretending to have too much a-priori economic theory. In *New methods in business cycle research*, ed. C. Sims et al. Minneapolis: Federal Reserve Bank of Minneapolis.

Singleton, K. 1980. A latent time series model of the cyclical behavior of interest rates. *International Economic Review* 21: 559–575.

Stock, J., and M. Watson. 1988. Testing for common trends. *Journal of the American Statistical Association* 83: 1097–1107.

Stock, J., and M. Watson. 2002. Macroeconomic forecasting using diffusion indexes. *Journal of Business and Economic Statistics* 20: 147–162.

Stone, J. 1945. The analysis of market demand. *Journal of the Royal Statistical Society, Series A* 108: 286–382.

Stone, J. 1947. On the interdependence of blocks of transactions. *Journal of the Royal Statistical Society, Series B*, Supplement, 1–45.

Vahid, F., and R. Engle. 1993. Common trends and common cycles. *Journal of Applied Econometrics* 8: 341–360.

Common Land

T. Williamson

The legal status of common land is a source of considerable popular confusion. With the notable exception of some village greens, commons do not represent areas which are owned by nobody, nor areas which are owned by everybody, nor even by everybody within a given locality. Since the early medieval period, commons have been owned by specific individuals, usually the lord of the manor within which they lie. The term 'common' refers not to ownership, but to rights held in common by certain people to use the product of the soil of the area in question. In turn, this means that the owner cannot enclose the land; hence the unfenced open space which is still the most characteristic feature of a common (Campbell and Clayden 1980).

These rights are not now, nor have they been in the historic past, generally shared by everyone living within a given locality. Instead they are usually attached to specific dwellings, or to their sites. Except, therefore, where the owner has so decreed, the use of commons as recreational open spaces by a wide public is not in the strict sense defensible in law.

Commons, of course, were not in origin primarily places for recreation. They formed a vital element in the pre-industrial rural economy. Commons represent the attenuated remnants of the medieval wastes: areas which were not used to produce arable crops, but to provide a range of other resources. The principle rights exercised by commoners were, and are: *common of pasture*, or the right to pasture animals; *pannage*, or the right to allow pigs to eat acorns or beech mast; *common in the soil*, or the right to take minerals, gravel, stone, sand etc. for use on the commoner's holding; *estovers*, or the right to take small branches, bracken etc. for fuel, fencing or animal litter; *turbary*, or the right to dig peat or turf for use as fuel; and *piscary*, the right to take fish from streams or ponds on the common. Of these rights, that of common of pasture has normally been by far the most important.

The early history of common land is obscure. It appears that in the early Saxon period, areas of open waste were much more extensive than they were to become in the medieval period, and the rights to their use were more loosely defined and often exercised by much wider groups. The name of Sherwood Forest, for example – the Shire Wood – indicates that it was once the common woodland of the entire shire of Nottingham. Limitation and closer definition of rights to common waste occurred during the population increase of the early medieval period (Hoskins and Stamp 1967). As arable expanded, common grazing dwindled, and areas of waste which had formerly been shared by communities were now divided between them, often after violent disputes. But there were also disputes within communities, as manorial lords, in association with their more prosperous tenants, attempted to take areas of the common waste into private ownership. The medieval struggle for the commons culminated in the Statute of Merton (The Commons Act of 1235), which decreed that freeholders had to be left with sufficient pasture to maintain the mixed farming of their holdings. However, the rights of the customary tenants were not protected by statute law, and the passing of the law for the first time clearly enshrined in national law the concept that the manorial lord rather than the community itself was the *owner* of the common waste of a manor.

The continual expansion of arable at the expense of the common wastes during the period before 1300 had other effects. There was an increasing tendency for commons to be *stinted*, that is, for the number of beasts put out by each commoner to be more carefully regulated. There were a number of ways in which this could be organised, the most usual being by the rules of levancy and couchancy, that is, where the right to turn out was measured by the capacity of the commonable tenement in such a way that only as many animals could be turned out as the tenement

was able to support (with the aid of hay etc.) through the winter.

The extent to which the commons survived during the medieval period varied considerably from region to region, as the result of the interplay of a number of factors. Essentially, commons survived best where population densities were low and where much land was unsuitable for arable agriculture. Thus large areas survived in the uplands of the north and west. But there may have been more complex social and economic factors which were also important in the preservation of common grazing, especially in lowland areas, for the distribution of commons in medieval England does not appear to be a direct and simple reflection of demographic pressure or environmental factors. In certain parts of the south and east of England – areas of dispersed settlement and irregular field systems, poorly developed communal controls and individualistic agriculture – some communities seem to have lacked the management structures necessary to act corporately to plough up areas of common grazing, even in areas of high population and moderately fertile soils, such as Norfolk. In addition, poor controls on the alienation of land and the practice of partible inheritance led to an early proliferation of smallholders for whom the resources provided by the commons were of vital importance. In such areas the conversion of waste to arable agriculture ground to a halt rather earlier than in the classic open-field areas of the Midlands.

In the latter areas, communities were more cohesive and communal controls on agriculture and land-use better developed. There were often stronger manorial controls on the alienation of land, holdings did not fragment to the same extent and there was less economic polarisation within the farming community. It may be this that explains why in many of these areas so much of the wastes were ploughed up in the 12th and 13th centuries, as arable prices, and the need for food for consumption, increased. Whatever the explanation, it appears that in many parts of the central Midlands, areas of common grazing were almost entirely destroyed by the end of the 13th century. The arable strips of many villages ran right up to the parish boundary, meeting with those of neighbouring villages.

This was not true of all areas in the Midlands, however. Conversion of grazing to arable was more retarded in the Forest areas. Such areas were not necessarily densely wooded; the term *forest* was a legal rather than a descriptive or environmental term, referring to areas to which forest law applied. This was a body of rules and restrictions originally intended to preserve deer for the royal chase and which *inter alia* attempted to limit the destruction of suitable habitats through the expansion of arable cultivation. In reality, forest law functioned more as a source of revenue, for encroachments on the wastes were tolerated if a fine was paid. Nevertheless, in areas like Rockingham Forest in Northamptonshire, these institutional factors combined with the relatively marginal nature of local soils to preserve extensive areas of open waste.

As arable expanded at the expense of grazing, in the areas of irregular field systems and dispersed settlement surviving commons often became foci for settlement. The poorly developed nature of communal controls in such areas was probably the stimulus for this development; farmers and smallholders moved to the edge of areas of common grazing not only for the convenience provided by such a location, but also, as it were, to stake a visible claim to their use. The freedom to alienate land and the irregular nature of field systems in these areas made such settlement migration possible, for they allowed the acquisition of blocks of land adjacent to the common upon which farmsteads could be established. This development seldom occurred in areas of regular open-field systems, nucleated settlement, and strong community controls.

Thus it was mainly in the south–east and the north–west of England that the medieval period saw the development of straggling settlements around the perimeter of commons. This process went furthest in parts of East Anglia, where complex manorial organisation and the presence of substantial numbers of free tenants practising partible inheritance led to a proliferation of smallholdings and wholesale migration away from earlier village sites. Farmsteads clustered around

areas of common which thus became large village greens, leaving the parish church – marking the original Saxon site of settlement – isolated in the fields some distance away (Wade Martins 1975).

Lowland commons have a distinctive form. They have straggling, concave outlines, formed by a series of rough arcs; roads funnel into the common where these arcs join. This characteristic shape probably derives from the fact that commons are the remnants of more extensive areas of waste which had been continuously encroached upon for centuries before their outline became fossilised, usually in the early medieval period. The perimeters of commons are often defined by particularly massive and ancient banks and ditches.

Today, many commons are wooded, but this is normally a relatively recent feature, resulting from a relatively recent decline in the intensity of grazing. By the end of their medieval period, commons had usually lost whatever woodland they had formerly carried; it had been destroyed by felling and over-grazing, and only names like 'Wood Green' sporadically reflect their former nature. Medieval woodland, in contrast, was not usually common land, but land which had been enclosed from the waste and over which the use and access of others had been limited (Rackham 1976).

By the end of the Middle Ages, there were considerable regional variations in the extent of common land. Commons survived better in the upland areas of the north and east than in the more fertile lowland zone. Within the lowland areas, they tended to survive better in the south and east of England, and in the west country, than in the classic open-field areas of the Midlands, with the exception of the forest areas, where they usually also survived well.

These variations were a factor in the local and regional development of rural society in the post-medieval period. In particular, areas in which extensive commons survived tended also to be the areas in which the decline of the small freehold farmer, which continued at varying rates throughout the post-medieval period, was retarded. In areas like the Fens, small farmers used their rights to extensive commons to adopt forms of livestock farming as specialized agricultural regions emerged in the 15th century. Survival of commons also allowed small cottagers to maintain a measure of economic independence, and the more extensive commons attracted large numbers of squatters, often part-time craftsmen. As a result, areas in which large commons survived tended to have a reputation for lawlessness. The opportunities which such areas offered to the poor also ensured that they often experienced particularly rapid population growth in the early modern period. In Northamptonshire, for example, forest villages in the 17th century were on average around half as populous again as non-forest villages (Hoskins and Stamp 1967, p. 52).

Much enclosure of common land occurred during the 16th and 17th centuries. Nevertheless, in 1688 Gregory King estimated that there were still 10 million acres of heaths, moors, mountains and barren land in England and Wales, and a further 3 million acres of forests, parks, and commons, the majority of which was common land. Today, the total area of common land in England and Wales is around 1.5 million acres. Even allowing for a high degree of inaccuracy in King's estimates, there has clearly been a dramatic reduction in the area of common land as a result of Parliamentary enclosures, mainly in the late 18th and early decades of the 19th centuries. Enclosure was principally inspired by a desire on the part of the larger landowners to profit from the conversion of the remaining commons to arable, or their improvement as pasture, both of which were difficult or impossible where the land was subject to the use and access of a large number of local inhabitants. The high prices of arable during the Napoleonic Wars were a particular stimulus to enclosure of open heaths and wastes, especially on the light soils of eastern England (Turner 1980, pp. 63–93).

The enclosure of commons was, like the enclosure of open fields, closely connected with the decline of the small owner-occupier which had continued almost uninterrupted throughout the early modern period. As land fell into the hands of relatively few people, so it became easier to obtain the agreement necessary to enclose, especially as with the advent of Parliamentary

Enclosure a majority in favour of enclosure was judged on the basis of the area of land which the agreeing landowners held, rather than on their number.

Yet as well as being in part a consequence of the decline of the small proprietor, enclosure of commons also served to accelerate this process. The allotments received by those small farmers or cottagers who were able to prove the legality of their claim to common rights were seldom sufficient compensation for the advantages lost through enclosure, especially when legal and fencing costs were taken into account. For many small farmers, enclosure was the final misfortune which led to their departure from farming; for the small cottager, enclosure often led to increased, if seasonal, reliance on poor relief (Snell 1985).

Today, the distribution of surviving common land in England and Wales continues to be very uneven, with more in the highland zone than in the lowlands. Within the lowlands, there are still fewer commons in the Midland counties than in the south and east, or in the west country. In lowland areas, most commons are now principally valued for their amenity value, or for their role as nature reserves or Sites of Special Scientific Interest. Common land now survives as such only where it has been registered under the terms of the Commons Registration Act of 1965. All existing common land is listed in the final register, which was closed on 1 August 1972.

See Also

- ▶ Common Property Rights
- ▶ Open Field System

Bibliography

Campbell, I., and P. Clayden. 1980. *The law of commons and village greens*. Henley-on-Thames: Open Spaces Society.

Hoskins, W.G., and L. Dudley Stamp. 1967. *The common lands of England and Wales*. London: Collins.

Rackham, O. 1976. *Trees and woodland in the British landscape*. London: Dent.

Snell, K. 1985. *Annals of the Labouring poor: Social change and agrarian England 1660–1900*. Cambridge: Cambridge University Press.

Turner, M.E. 1980. *English parliamentary enclosure*. Folkestone: William Dawson.

Wade Martins, P. 1975. The origins of rural settlement in East Anglia. In *Recent work in rural archaeology*, ed. P.J. Fowler. Bradford on Avon: Moonraker Press.

Common Law

P. S. Atiyah

Common law is a system of law and legal processes which originated in England shortly after the Norman Conquest and after several centuries of continuous development was exported to the English colonies, and so came to be the basis of the law of the greater part of the United States, as well as of Australia, New Zealand, most of Canada and (to a lesser degree) also of India, Pakistan, Bangla Desh and many parts of Africa. The chief characteristic of the common law has always been that its development has lain largely in the hands of the judges, and that it has therefore grown and changed incrementally, case by case, in the course of actual litigation.

In modern times the term ‘common law’ is used in a variety of senses. In the broadest sense, it continues to be used to refer to the entire system of law originating in England which now forms the basis of the law in the greater part of the former British Empire, often nowadays called the ‘common law world’. In this sense the common law is often contrasted with the ‘civil law’ which derives from the law of ancient Rome, and today operates in most of Western Europe, as well as in a number of other countries (such as Japan and Egypt) which have borrowed their law from European countries. One of the chief characteristics of the modern civil law is that it derives its authority from one or more basic Codes of law; and it remains a principal distinction between common law and civil law countries that the former have not generally codified their law. And even in

common law jurisdictions (such as California, for example) where there does today exist a kind of common law Code, it differs fundamentally in nature from the civil law Codes; in particular the system of precedent, and the authority of the judges to interpret and develop such common law Codes are quite different from those recognised in civil law countries.

The term 'common law' is also often used in various narrower senses. In the most important of these narrower senses, the common law is often contrasted with legislation, so that the lawyer in a common-law country still thinks of legislation as a type of law different from the 'common law', which is basically judge-made law. The term 'common law' is sometimes used in yet a third relevant sense in which it is distinguished from a body of law, known technically as 'Equity' which was originally supplementary to the common law, and was developed in the separate Court of Chancery. Today common law (in this narrow sense) and 'Equity' are almost everywhere merged and administered by a single set of courts.

The common law (in the first two senses identified above) has traditionally been associated with the economics of the free market in at least two different ways. First, there is a strain of thought, represented in particular by Hayek (1973), which seems to suggest that a system of law, like the common law, which is largely judge-made, is inherently more likely to favour and protect individual freedoms, and among them (or especially) economic freedoms. But this is an implausible and indeed eccentric claim, which seems to involve confusion of the first two senses of the term of 'common law' referred to above. Because most redistribution is accomplished in modern democracies by legislative measures, it is easy to assume that a legal system which owes little to legislation will be more likely to recognize and protect the freedom of the market, but the amount of redistribution which occurs in a legal system does not necessarily depend upon whether that society is part of the common law world. There is no a priori reason to suppose that judges left to themselves by a legislature will necessarily favour the economics of the market. In the last analysis, the policies favoured by judges will depend upon their own preferences, their culture and traditions.

But there is a second way in which the common law has traditionally been associated with the freedom of the market, and this association rests upon the historical facts of the last three centuries. The concept of the Rule of Law which came to be recognized and defended in England after the revolution of 1688 has been seen by many as having favoured the development of a free market economy in England prior to and during the early years of the industrial revolution. Because of the historical fact it was for a long time almost an article of faith among English writers that the common law and the freedom of the market were closely associated. This view is today less strongly held in England, as a result no doubt of the fact that, while Englishmen still like to believe in the Rule of Law (despite grave doubts in some quarters as to whether this concept has much meaning), they are by no means so wedded to the free market as they were. In America, where the Constitution of 1788 substantially embodied the English traditions as to the Rule of Law, as well as the then accepted ideology of the free market, the association between the two has survived rather more strongly.

The reasons for the traditional belief in the close association between the common law and the freedom of the market must therefore be sought in history, and in particular in English history during the period from approximately 1770 to 1870, when the free market economy was largely in process of being established. And of all parts of the common law, none was more important for this purpose than the law of contract, because this was the part of the law most intimately related to the economic system. Indeed, the story of English law between 1770 and 1870 was to a large degree the story of how the law of contract was converted into the law of the free market, and of how the ideology of freedom of contract became one of the great intellectual movements of history (Atiyah 1979).

The first three-quarters of the 18th century was a period of transition in England, during which many older ideas about contract and the market were being displaced by the newer ideas which

gradually became dominant towards the end of the century. Among the older ideas at least three can be identified as particularly hostile to the laws needed to serve the emerging free market economy. First, there was a regulatory element in the law and the economy dating back to Tudor times, represented for instance by statutory controls of wages and prices of many commodities, and by the apprenticeship laws which controlled entry to many trades with outdated and largely unnecessary restrictions. Secondly, there was a paternalistic element in much contract law at this time, with the courts still being willing to relieve various classes of persons from the consequences of bad bargains which they had made. This paternalism was particularly pronounced under various doctrines of Equity, such as rules for the relief of mortgagors, rules against the enforcement of contractual penalties and forfeitures, rules for the protection of seamen and 'expectant heirs', and so forth. Thus, in the third sense of the term the 'common law' identified above, it can be said that the common law was always more market-oriented than Equity. Thirdly, there was a traditional moralistic element in the contract law of the 18th century, and this also took different forms, such as the general hostility to usury (as to which see Simpson 1975, pp. 510–18), and the attempts to regulate the way in which essential foods and drinks were sold by use of the traditional marketing offences. The 'moral' roots of older law were also related to ideas about 'just prices' which, though rarely openly recognized in the common law, seem to have been influential at least in some of the cases in Equity, where there are signs that the Chancellors did have some vague sense of unease if they were asked to enforce contracts at prices which seemed to them very unfair, or on terms which were (in the language of the law) 'unconscionable'.

In addition to these specific instances of interference with the binding force of private contracts, there were important respects in which the whole concept of a general contract law remained relatively undeveloped at this time. Thus, while the law recognized and enforced specific types of contracts, such as contracts for the sale of land, contracts of insurance and so forth, there was, as yet, little sign of a general law of contract, governing all types of transaction. Then also, it remains unclear how far the contract law of this period actually recognized and enforced wholly executory contracts, in the sense of awarding damages for breach of a contract prior to any acts of performance or detrimental reliance by any of the parties. And finally, it is clear that, from the standpoint of today, the law of contract in the 18th century had not yet freed itself from dependence on the law of property. Of course, in one sense contract law can never be free from a dependence on property entitlements, because contract law is the mechanism by which entitlements are exchanged; but there are clear signs in the 18th century that contract law was still closely tied to property law in another sense, in the sense (for instance) that the proprietary aspects of many transactions were still regarded as more important than the promissory or contractual aspects. So, for instance, the right of a mortgagor to redeem the mortgaged property was protected by the courts, even when by the terms of the mortgage documents he had forfeited that right by delay in repaying the loan. It was assumed that if the mortgagee received back his money, with interest and costs, he was adequately protected by the law, even though the contract itself would have given him more extensive rights.

During the century beginning around 1770 these older ideas and traditions gradually gave way before the ideology of freedom of contract; but it would be wrong to think that this ideology did not have long roots and antecedents in still earlier periods. There are, even in the 16th and 17th centuries, many signs of incipient economic liberalism among the lawyers such as Coke, who bequeathed to the common law a hatred of monopolies as well as a passion for individual liberties (Wagner 1935). And Thomas Hobbes, in a well-known passage in *Leviathan*, had swept away all the medieval learning about 'just prices' and declared that '[t]he value of all things contracted for, is measured by the Appetite of the Contractors; and therefore the just value, is that which they be contended to give' (Hobbes [1651] 1968, p. 208). So the ideology of freedom of contract certainly had origins going back well

beyond the 18th century. Nevertheless, it does seem (though the matter remains controversial) that major changes in the law began during the course of that century which gathered pace as the century progressed.

Certainly, a great deal occurred to change the character of contract law from the last quarter of the 18th century until well into the 19th century, and there is much evidence that many of these changes in the law were profoundly influenced by classical economic theory, and perhaps still more by popular versions of classical economic theory. First, the relics of the Tudor regulatory economy gradually disappeared. Wage regulation had become increasingly obsolete in practice during the 18th century, and a major challenge to the older laws in the name of freedom of contract had taken place in the celebrated case of the Gloucestershire Weavers (1756–7), (Atiyah 1979, pp. 73–4). By the early 19th century most of the legislation authorizing the fixing of wages had been repealed. So too was the Statute of Apprentices, after many years during which its operation had been gradually whittled down by the judges. Secondly, the signs of paternalism which are still found in 18th-century Equity seem to have disappeared gradually as the judges hardened their hearts and toughened their minds. For example, signs of an attempt to introduce implied warranties on the sale of goods for the protection of buyers, which can be detected in the 18th century, were largely scotched, and the principle of *caveat emptor* reasserted with full vigour. The equitable doctrines allowing the courts to relieve various unfortunates from the effects of hard bargains were gradually whittled down, although they never disappeared altogether. Third, the moralistic elements in the law were also gradually whittled down. The law of contract came increasingly to be seen to be neutrally enforcing agreements which must be presumed to be beneficial to both parties. The only moral component left in the law of contract during the 19th century seemed to derive from the binding nature of promises.

The subjective theory of value also seems to have been largely accepted by the judges even before it had been wholly accepted by economists. Although the common law had always insisted that a promise be supported by some 'consideration', some reason, before it would be enforced (and to that extent at least contained a paternalist element), the growing acceptance of the subjective theory of value meant that the doctrine of consideration became much less important during the 19th century. So far instance, in *Haigh* v. *Brooks* (1840, 113 English Reports 124) the judges enforced a promise to pay £9000 in return for the giving up of a guarantee previously given by the promisor, even though it now appeared that the guarantee might be unenforceable and legally worthless. The promisor had valued it at £9000, said the judges; it was not for them to say that the document was worthless. For similar reasons, the prejudice against usury had gradually been overcome, and the usury laws were totally repealed in England in 1854.

In these ways, then, the principle that contracts are binding and must be strictly enforced had been greatly strengthened, and exceptional cases had been whittled down by the middle of the 19th century. In addition, other changes had occurred in the general nature of contract law, which were closely related to the growing trend to see contract law as the law of the free market. First, it was during this period that a general law of contract came into existence for the first time in the common law world. And the process of generalization was important to the ideology of the law in a number of respects. In particular, the generalizing of contractual ideas meant that the law had to become more abstract, more broadly principled. Principles had to be developed which could be applied equally to (say) commercial contracts for the sale of wheat, to contracts of employment, and (for instance) to personal contracts such as the contract to marry. This abstraction may have helped the law become more neutral, less inclined to pursue any redistributive tendencies, such as may exist where (say) there is a separate body of legal doctrine dealing with contracts of employment, or with residential leases, or with loan transactions.

Next, it seems clear that another major development during this period was the gradual shift in emphasis in contract law away from treating

contracts as present, or partly performed exchanges, and towards treating them as private planning devices, made in advance to allocate risks. The wholly executory contract became clearly recognised by the law, so that it now became possible for a person to sue for damages for breach of a pure promise, even where no performance or detrimental reliance had taken place. The justification for requiring damages to be paid in such circumstances was never clearly enunciated, and indeed, specific justification was rarely seen to be necessary. It was widely assumed that the broad principle of freedom of contract required, not only that parties be left free to make their own exchanges, but that the law should be available in aid of a party to enforce his claim to damages where the other failed to perform. John Stuart Mill was the first economist to point out that a policy of laissez-faire could not be used to justify the enforcement of executory contracts (Mill 1848, vol. 2, p. 386), but even modern economists do not generally pursue this line of thought, though some libertarians have done so.

And finally, 19th-century contract law increasingly freed itself from its dependence on property law. Although obviously entitlements still remain the subject matter of all contracts, contract law has become much less concerned with specific items of property, and is more concerned with wealth as a kind of fungible property. The reason for this was basically that 19th-century contract law was dominated by the needs of merchants and traders, to whom all property is in principle replaceable with money. A merchant can be assumed to be indifferent between a piece of property, and the value of that property. Similarly, as contracts came to be increasingly seen as fundamentally risk-allocation devices, the particular entitlements or property to which the risks attached became less important.

By the last quarter of the 19th century, the process of developing a mature body of general contract law had largely been completed in England, and although a similar process took place in America (Horwitz 1977), there is ground for believing that that was not completed for another fifty years or so. Freedom of contract had, apparently, reached its highest point. But although this was true of the ideology of freedom of contract among lawyers and judges, it was not really true of the views of economists or of the politicians, or of the public. By the late 19th century, neoclassical economists were already beginning to write sceptically about the sweeping effects of freedom of contract which had been attributed to the classical economists, and were pointing out the many possible causes of market failure such as information difficulties, externalities and monopoly. And although most of the older regulatory legislation had been repealed in the first half of the 19th century, Parliament had at the same time been gradually building up a completely new body of regulatory enactments dealing with new industrial problems – factories, coal mines, safety at sea for seamen and emigrant passengers, public health, the adulteration of food and drink, regulation of the weights and measures used for sales, and so on. Much of this new legislation had been a pragmatic response to perceived evils, and though some of it could have been justified economically by arguments concerning misinformation or externalities, much of it would have been difficult to justify except on the assumption of paternalistic or redistributive motives. Some of it may have been inspired by sheer impatience, an unwillingness to give the market time to work, or a belief that the short-term costs of market failures were so severe that legislative correction was necessary without regard to the long term distortions this might produce.

What is quite clear is that by the time the English common law and common lawyers had accepted the teachings (as they were thought to be) of the classical economists on freedom of contract, these teachings were already somewhat out of date. The result was that the mature common law of contract was seriously deficient in a number of respects. It was first of all deficient in its almost total neglect of the problem of externalities. Contracting parties were entitled to pursue their own interests, regardless of the effect of their

contract on third parties, or the public. Only in the most extreme cases of actual illegality would the courts generally refuse to uphold a contract. Secondly (although this certainly could not be laid at the door of the classical economists), there had been, during the 19th century, a serious neglect by common lawyers of the problem of monopoly. This may well have been largely due to the fact that for the greater part of this period the British economy was itself highly competitive, and in little danger from monopolies. But the complacent assumption that cartels were unstable and were always vulnerable to internal or external competition was in England (though not in America) carried over by lawyers and courts into new conditions towards the end of the nineteenth century, and well into the present century, when it was utterly out of date. A second result of this failure of the common law to keep pace with economic theory and political reality, was the growing gulf between the common law and legislation. Once again, extensive legislative intervention with freedom of contract began to become commonplace, and much of it was increasingly redistributive in character.

During the course of the present century this process continued at an increasing pace until 1980 or thereabouts, since when there are signs that history has virtually reversed itself. Disillusion with the free market, particularly in England, increased during the great depression in the 1930s until, by the end of World War II, a Labour Government was elected to power with a massive majority and with a mandate to lay the foundations for a socialist state and a socialist economic system. Since then England has increasingly learned to live with a 'mixed economy', to a large part of which the traditional law of contract seems irrelevant because the public sector is often controlled by public laws rather than by contract law. But even in areas where private law continues to operate, the common law of contract has become increasingly affected by legislative intervention. Virtually all types of consumer transactions are today controlled or affected to some degree by legislation, including consumer credit contracts, contracts of employment, residential leases, and insurance contracts. Unconscionable, or unfair contracts are increasingly subjected to judicial control. Many areas of law which were formerly controlled largely by contract, such as family law, are now subject to extensive judicial discretionary control. Even business and commercial contracts are subject to vast bodies of legislative and regulatory laws, some, such as the modern monopoly anti-trust laws, being designed to preserve the operation of a competitive market, but much of it still being designed to restrict competition or the operation of the free market.

America has not gone so far down this road as Britain and other common-law countries, and indeed, for a long time, in the late 19th and early 20th centuries, constitutional decisions of the United States Supreme Court in the name of freedom of contract, actually prevented similar developments. Much legislative intervention with freedom of contract was, during this period, declared unconstitutional, frequently over the dissent of Justice Holmes. By the late 1930s, however, the majority of the court had largely accepted Holmes's view, and since then, legislative intervention with freedom of contract has not been regarded as per se unconstitutional. This shift in the court opened the door to the same kind of regulation and intervention which had already been taking place in Britain, and although America has not, like Britain, brought large-scale industries within the public sector and therefore partially outside the control of contract law, most of the other legislative developments of the British type certainly have their parallel in America. No doubt some contracts are more regulated in Britain, but conversely there are plenty of examples of legislative interference with freedom of contract in America which are not to be found in Britain.

These vast changes in the operation of the common law have accompanied or brought with them a change in ideology once again. Paternalism and redistribution were, at least until around 1980, increasingly favoured by many writers and

teachers of contract law, as well as large sectors of the electorate. Even the judges became much more sympathetic to arguments based on concepts like unconscionability and inequality of bargaining power. In America, unconscionability was given express legitimacy as a device for overturning unfair contracts by the Uniform Commercial Code, and was increasingly used by the judges as a matter of common law as well. Many relationships of a contractual character (for instance, that of physician and patient) and others of a virtually contractual character (for instance, that between manufactures of products and ultimate purchasers and consumers) are, both in American and Britain, increasingly regulated by tort law rather than contract law, at least where things go badly wrong and legal actions for damages are brought based on negligent conduct, or on defects in the goods. In such malpractice or products liability actions the appropriate standards of care or quality are set by judges and juries and not by the contracting parties, and contractual exculpatory clauses are often denied legal validity.

Since about 1980 there have been increasing signs that the tide has turned yet again, both in Britain and America. Obviously, and visibly, British and American governments have since then been trying to reassert the virtues of the free market and roll back the frontiers of regulation, and in this they are being vigorously supported by some lawyers and law teachers in America, though not to any real extent in Britain. It is not yet clear what the impact of this is going to be on the future of the common law of contract. One possible scenario is that, as in the late 19th century, the courts will be behind the times, but that on this occasion they will be hostile to the reasserted belief in the free market and will continue to defend paternalist and redistributive intervention in free contracts, particularly where one of the parties to the contract is a consumer or 'small man' thought to be weak in bargaining power. But another possible scenario is that the new enthusiasm for the free market will prove but a short-lived hiccup in the long-term trend towards paternalist and redistributive policies. In either event it seems unlikely that for many years to come British or American courts will be enforcing contracts according to the full rigour of the common law.

See Also

► Law and Economics

Bibliography

Atiyah, P.S. 1979. *The rise and fall of freedom of contract*. Oxford: Oxford University Press.

Hayek, F.A. 1973. *Law, legislation and liberty*, Rules and Orders, vol. 1. London: Routledge & Kegan Paul.

Hobbes, T. 1651. In *Leviathan*, ed. C.B. Macpherson. Harmondsworth: Penguin Books. 1968.

Horwitz, M.J. 1977. *The transformation of American Law 1780–1860*. Cambridge, MA/London: Harvard University Press.

Mill, J.S. 1848. *Principles of political economy.* From the fifth London edition. New York: D. Appleton & Co., 1908.

Simpson, A.W.B. 1975. *A history of the common law of contract*. Oxford: Clarendon.

Wagner, D.O. 1935. Coke and the rise of economic liberalism. *Economic History Review* 6(1): 30.

Common Property Resources

Jean-Philippe Platteau

Keywords

Common property resources; Efficiency; Market integration; Open access; Private property rights; Tragedy of the commons

JEL Classifications

O1

The concept of common property has become famous in economics since Garett Hardin (1968) wrote his celebrated article on 'The Tragedy of the

Commons'. In this article, common property is taken to mean the absence of property rights in a resource, or what is equivalently known as a regime of 'open access'. Under such a regime, where a right of inclusion is granted to anyone who wants to use the resource, Hardin argued, inefficiency inevitably arises in the form of over-exploitation of the resource accompanied by an over-application of the variable inputs. Open access leads to efficiency losses because 'the *average product* of the variable input, not its *marginal product*, is equated to the input's rental rate when access is free and the number of exploiters is large' (Cornes and Sandler 1983, p. 787). The root of the problem lies in the fact that the average product rule does not enable the users to internalize the external cost which their decisions impose on the users already operating in the resource domain. Of course, the efficiency losses are conceivable only in a world of resource scarcity, implying that the variable input is subject to decreasing returns. Such losses are considerable since they amount to the dissipation of the whole resource rent. Here is the crucial intuition behind the open access regime: when no property right is attached to a resource, the value of this resource is zero in spite of its scarcity.

Efficiency losses are to be measured not only in static but also in dynamic terms. Indeed, in an open access regime resource users are induced to compare average *instantaneous* returns with the input's rental price even though they may well be aware that they thereby contribute to reducing the future stock of the resource. The problem is simply that they are forced to follow a myopic rule because there is no way in which they can reap the future benefits of restraint in the present. Thus, for example, by refraining today from catching juvenile fish or from cutting down saplings in the forest, a villager can receive no assurance that he or she will be able in the next period to catch mature fish or to fell fully grown trees.

The main criticism levelled by numerous social scientists against the concept of open access is that the corresponding regime is rarely encountered on the ground. The typical regime, according to these critiques, is one under which a community possesses a collective ownership right over local natural resources. Under common property, therefore, a right of exclusion is assigned to a well-defined user group, and Hardin has created a lot of confusion by using the word 'commons' to refer to the alternative situation where no such right is granted to any agency. What is not always clear, however, is whether the ownership right involves only the ability to specify the rightful claimants to the resource, or whether it also involves the ability to define and enforce rules of use regarding that resource (for example, regulations about the harvesting season and production tools, allowed quotas of harvestable products of the resource, or taxes). Baland and Platteau (1996) have coined the term 'unregulated common property' to refer to the former situation, while the term 'regulated common property' is used for the latter.

Two polar situations can be considered on the basis of this analytically important distinction between two types of common property regimes. At one extreme, if common property is perfectly regulated, in the sense that the rules of use designed and enforced by the owner community allow a perfect internalization of the externalities, common property becomes equivalent to private property with a sole owner from an efficiency standpoint. This illustrates the general result that, absent transaction costs, institutions do not matter. At the other extreme, a strictly unregulated common property in the above sense implies that, as the number of users becomes quite large, over-exploitation of the resource becomes as important as under the open access regime: the rent attached to the resource is totally dissipated (see Platteau 2000, ch. 3).

Between these two extremes we find the situations most typically observed on the ground and described in the numerous field studies devoted to this topic (see Ostrom 1990; Baland and Platteau 1996, for a review of such studies). In such instances, rules of use exist alongside membership rules, yet they tend to be imperfectly designed and imperfectly enforced by the village community. One key reason for these imperfections is the governance costs that unavoidably plague any collective decision-making process. Governance

costs include all those costs incurred to reach a collective agreement and to organize a community of users. They are likely to be higher when the group is larger and when its membership is more heterogeneous (whether measured in terms of diversity of objectives or of wealth inequality). Moreover, governance costs are enhanced by the opportunistic tendencies of rights-holders not only to violate or circumvent collective rules but also to eschew efforts to create collective mechanisms of decision-making and enforcement. Costs arising from these proclivities are also dependent on the size of the user group: they are lower if the number of resource users is smaller and, at the limit, they are nil when there is a single user.

As a consequence of the aforementioned limitations, resources are less efficiently managed under a common property regime than they could be under a private ownership system. This is especially true if, owing to their scarcity, the resources carry high values which should be reflected in high rents. Population growth and market integration are thus two forces that tend to increase the monetary value of the efficiency losses arising from common property, that is, the forgone rents. This, at least, is the conclusion drawn by the so-called property rights school of Chicago economists (see, for example, Demsetz 1967; Barzel 1989). The advantages of private property appear all the more decisive as such a regime enables users to internalize externalities without incurring any governance costs. This is because it establishes a one-to-one relationship between individual actions and all their effects: 'A primary function of property rights is that of guiding incentives to achieve a greater internalization of externalities ...' (Demsetz 1967, p. 348).

Nevertheless, this ignores the costs of privatizing natural resources, which involve both directs costs and opportunity costs. Direct costs comprise transaction costs, such as the costs of negotiating, defining and enforcing private property rights. The usual argument is that such costs increase with the physical base of the resource. Thus, the wider the resource base (or the less concentrated the resource) the higher are the costs of delimiting and defending the resource 'territory' (Dasgupta 1993, pp. 288–9). For many natural resources, the costs of dividing the resource domain appear prohibitive under the present state of technology. For example, the open sea – or, more exactly, the fish stock contained in it – presents insuperable difficulties for private appropriation. The enforcement of exclusive property rights to individual patches of the ocean would, indeed, be infinitely costly. This is especially evident when fish species are mobile and move within wide water spaces, since exclusive rights are too costly to establish and enforce whether over the resource or over the territory in which the resource moves.

The opportunity costs of privatization, for their part, correspond to the benefits that are lost when the common property regime is abandoned. Here, we can think of scale economies that may be present not only in the resource itself but also in complementary factors. The obvious advantage of coordinating the herding of animals so as to economize on shepherd labour in extensive grazing activities is probably the best illustration of the way scale economies in a complementary factor may prevent the division of a resource domain. Another important category of opportunity costs is the insurance benefits associated with common property. When returns to a resource are highly variable across time and space, the need to insure against such variability is yet another consideration that may militate against resource division. When a resource has a low predictability (that is, when the variance in its value per unit of time per unit area is high), users are generally reluctant to divide it into smaller portions because they would thereby lose the insurance benefits provided by keeping the resource whole.

For instance, herders (fishermen) may need to have access to a wide portfolio of pasture lands (fishing spots) in so far as, at any given time, wide spatial variations in yields result from climatic or other environmental factors. On the assumption that the probability distributions are not correlated too much across spatial groupings of land or water

and that they are not overly correlated over time, a system offering access to a large area within which right-holding users can freely move appears highly desirable from a risk-reducing perspective.

The conclusion of the above discussion is, therefore, that the balance of the advantages and disadvantages of various property regimes is a priori undetermined. Economic theory, however, does provide useful guidance about which circumstances are more favourable to the persistence of common property or, conversely, to its demise and replacement by private property. Furthermore, instead of being fixed once for all, the balance sheet is susceptible to evolution depending on the transformation of the parameters on which the benefits and costs of privatization depend. Thus, the direct costs of resource division may fall with technological progress. For example, the introduction of modern borehole drilling facilitates the privatization of common grazing areas (Peters 1994). It is therefore not only the factors which enhance resource value but also those which reduce the direct costs of partitioning that may favour the private appropriation of natural resources.

See Also

- ▶ Access to Land and Development
- ▶ Agriculture and Economic Development
- ▶ Land Markets
- ▶ Population and Agricultural Growth
- ▶ Property Law, Economics and
- ▶ Tragedy of the Commons

Bibliography

Baland, J., and J.-P. Platteau. 1996. *Halting degradation of natural resources: Is there a role for rural communities?* Oxford: Clarendon Press.

Barzel, Y. 1989. *Economic analysis of property rights*. Cambridge: Cambridge University Press.

Cornes, R., and T. Sandler. 1983. On commons and tragedies. *American Economic Review* 83: 787–792.

Dasgupta, P. 1993. *An inquiry into well-being and destitution*. Oxford: Clarendon Press.

Demsetz, H. 1967. Toward a theory of property rights. *American Economic Review* 57: 347–359.

Hardin, G. 1968. The tragedy of the commons. *Science* 162: 1243–1248.

Ostrom, E. 1990. *Governing the commons: The evolution of institutions for collective action*. Cambridge: Cambridge University Press.

Peters, P. 1994. *Dividing the commons: Politics, policy, and culture in Botswana*. Charlottesville/London: University Press of Virginia.

Platteau, J.-P. 2000. *Institutions, social norms and economic development*. London: Harwood Academic Publishers.

C

Common Property Rights

Steven N. S. Cheung

In a society where individuals compete for the use of scarce resources, some rules or criteria of competition must exist to resolve the conflict. These rules, known as property rights, may be established in law, in regulation, in custom or in hierarchy ranking. The structures of rights may take a variety of forms, ranging from private property rights at one extreme to common property rights at the other. Most fall somewhere in between: either set of rights would be rare in its purest form.

In a private property, the delimitation of the right to its use is expressed in dimensions or characteristics inherent in the property itself. These rights are exclusive to some private party, are freely transferable, and the income derived from them is not attenuated, restrained or infringed by laws or regulations. Hence price control, taxation, and social restriction of transferability may be regarded as violations of private property rights. In a common property, there is no delimitation or delineation of its use rights to any private party. No one has the right to exclude others from using it, and all are free to compete for its use. Hence there are no exclusive use rights, no rights to be transferred, and in the limiting case, no

net income can be derived from using the common property.

This last condition rests on an economic proposition known as the dissipation of rent. It argues that because of the lack of exclusive use rights, individuals competing for the use of a common property will reduce its rental value or net worth to zero. The reason is that if no one has an exclusive claim to the value (i.e. rent) of that property, its use will invite competition to the point that each and every competing user can earn no more than the alternative earning of his own resources required in the exploitation of that common property. In other words, under competition and with no one having a special advantage, a 'prize' that has no exclusive claimant will be dissipated or absorbed by the costs of other resources which must be dedicated to its winning. Hence the net value of the prize won is zero.

The usual examples of common property rights cite a public beach and marine fisheries, and the dissipation of rent typically implies excessive use or over-exploitation. However, the dissipation may take the form of under-exploitation. For example, a piece of fertile land under common ownership may be used for herd grazing, or left idle, instead of being planted as an orchard.

In the real world, the complete dissipation of rent is rare indeed. This is because the supply curve of labour or of other inputs may be rising (some intramarginal rent may be captured), the competing users may have different opportunity costs (the non-marginal users may be enjoying rent), or entry may be restricted by regulations, by customs or by information costs. Still, with common property rights some dissipation of rent is inevitable, and no society can afford to surrender a large portion of its valuable resources to this structure of property rights.

A property may be held in common because its capturable rent is lower than the cost of enforcing exclusivity. In this case, the dissipation of rent is no waste. However, to the extent that rent dissipation is viewed as a waste, its occurrence must be attributable to the omission of some constraints in the analysis. Attempts to reduce rent dissipation go far to explain why common property in its 'pure' form is seldom observed. In marine fisheries, for example, numerous regulations govern the fishing season, the size of fish caught, the boat size and the mesh size, and various licensing arrangements restrict the number of boats and fishermen. The market value of a fishing licence, sometimes enormous, is one measure of the ocean rent captured. Even for public beaches, regulations of some type will often be found to govern the use of those most in demand.

Whereas regulations and restrictions on entry in the use of a common property often serve to reduce dissipation, the rent that can be captured is usually less than if the property were privately owned. To reconcile this observation with constrained maximization, we must infer that, enforcement costs aside, other transaction costs associated with the changing of institutional arrangements must restrain the formation of private property rights.

No economy can survive if the majority of its scarce resources are commonly owned. Regulations may, indeed, reduce rent dissipation; but in the process they not only distort the use of the resources but also invite corruption and the emergence of special interests. An unrestrained common property, strictly speaking, is propertyless in ownership; if its structure is extended to all resources, starvation for all must result. If one rules out private property rights, then to avoid the imposition of an infinite array of regulations the remaining alternative is the communal system or the communist state.

In a communist state there is no private owner of productive resources: each constituent is propertyless, in the literal sense of the word. Since the dissipation of rents associated with common properties will guarantee starvation, in a communist state the rights to use resources, and to derive income therefrom, are defined in terms of rank. That is, stripped of all ownership rights over valuable and productive resources, the citizens of a communist state hold differing rights to use resources and to obtain income according to

their status. In the people's communes in China under the Great Leap Forward, for example, no one owns the productive resources (i.e. everyone is propertyless), but comrades of different ranks enjoy different rights and privileges. 'Rank' as such has value and is subject to competition, therefore a system of 'property' rights is implicit. However, the valuable rights are now defined in terms other than the inherent properties of the productive resources.

This is, in fact, the key distinction between a private property system and a communist state: the former delineates rights in terms of certain dimensions of the productive resources themselves; the latter delineates rights in terms of a characteristic (rank) of people deprived of productive human capital. In the communist state, the competition for and protection of rank will draw on the use of valuable and productive resources (another form of rent dissipation). Moreover, the lack of market prices increases the cost of information, and the lack of contractual choices increases the cost of enforcing performance. What is saved in return are the costs of delineating and enforcing rights in properties.

It is among these varied costs – broadly defined as transaction costs – that we find the key divergence in economic performance between the communist and the private property systems. If one ignores transaction costs, the delineation of rights in terms of rank will produce the same use of resources as would the delineation of rights in properties. However, it can be convincingly argued that the broadly defined transaction costs are generally higher with communal than with private rights. Communism fails, not because it does not work in theory, but precisely because in practice its costs of transaction are higher than those in a system of private property rights. Still, the delineation of rights in ranks is a way to reduce rent dissipation in a propertyless state.

Strictly speaking, the dissipation of rent associated with common property is no 'theory' at all, because dissipating rent merely to produce an equilibrium does not explain behaviour. Worse, to stand aside and simply permit rent to dissipate is inconsistent with the postulate of constrained maximization.

What is useful and important from the standpoint of economic explanation is to view whatever rent dissipation does occur as necessarily a constrained minimum because, under the maximization postulate, each and every individual has an incentive to reduce that dissipation. Behaviour associated with the dissipation of rent must therefore be regarded as attempts to reduce that loss, and this altered view explains many observations. That some dissipation remains must then be attributable to the constraints of transaction costs. The challenge to the economist is to specify and identify what these costs are and how they will vary under differing circumstances.

See Also

- ▶ Coase Theorem
- ▶ Fisheries

References

Alchian, A.A. 1965. Some economics of property rights. *Il Politico* 30(4): 816–829.

Bottomley, A. 1963. The effect of the common ownership of land upon resource allocation in Tripolitania. *Land Economics* 39: 91–95.

Cheung, S.N.S. 1970. The structure of a contract and the theory of a non-exclusive resource. *Journal of Law and Economics* 13: 49–70.

Cheung, S.N.S. 1974. A theory of price control. *Journal of Law and Economics* 17: 53–71.

Cheung, S.N.S. 1982. *Will China go 'capitalist'?* Hobart paper 94. London: IEA.

Coase, R.H. 1960. The problem of social cost. *Journal of Law and Economics* 3: 1–44.

Demsetz, H. 1964. The exchange and enforcement of property rights. *Journal of Law and Economics* 7: 11–26.

Gordon, H.S. 1954. The economic theory of a common property resource: The fishery. *Journal of Political Economy* 62: 124–142.

Knight, F.H. 1924. Some fallacies in the interpretation of social cost. *Quarterly Journal of Economics* 38: 582–606.

Common Rights in Land

Leigh Shaw-Taylor

Abstract

'Common rights' and 'common land' refer to rights to use land in common in some way. Of several forms of common rights in pre-industrial Europe and elsewhere, only one – free access to land – involved what economists commonly think of as common rights. Common rights in Europe were largely swept away during the 18th and 19th centuries by a process termed 'enclosure'. Some economic historians have reconsidered the inefficiency of open fields in an English context, but at present the data are too poor to allow a plausible rebuttal of the views of 18th-century critics of the open fields.

Keywords

Access to land; Common fields; Common land; Common property resources; Common rights; Common rights in land; Common-pool resources; Enclosure; Open-access resources; Tragedy of the commons

JEL Classifications

N5

Common rights are rights to use land in common. The most important of these rights was the right to graze livestock on common grassland. But rights to gather fuel (wood, peat, gorse and turves), fertilizers, timber for building and other natural resources were also important. Common land is land used by a number of distinct individuals or households whose rights over the land are known as common rights.

Today we are accustomed to think of land as private property with a clear owner and possibly a tenant. Although in some countries there may be legal rights of public access to certain types of wild or agricultural land, it is generally the case that the owner or tenant of the land has exclusive rights to use the land and, within the limits of planning or zoning laws, may use it as he or she wishes. But in Europe, for at least a thousand years and ending only in the 19th century, a high proportion of land was 'common land' which many individuals were entitled to use for a variety of purposes.

It cannot be overemphasized that common land was generally not open-access land – land which anyone could use. There were regulations governing who could use the land, what they could use it for and how much they could use it. When economists think of common land and common rights they may have Garett Hardin's 'tragedy of the commons in mind' (Hardin 1968). The principal subject of Hardin's article was in fact population growth, not historical common land or common rights.

However, Hardin used a theoretical common land system as a model for the exploitation of open-access resources. In this system each herder could put as many animals as he wished on to the common pastures. Hardin argued that individual herders would choose to graze more and more animals on the common, thus inevitably leading to over-grazing and degradation of the resource. This model offers important insights into the destruction of, or damage to, unregulated open-access resources such as the atmosphere or fish stocks in the oceans. If common land and common rights had operated in this manner, it is unlikely that they would have remained a key part of European agriculture for so many centuries.

In the rest of this article the following questions are addressed: what were common rights? What was common land? Who had common rights? How was common land regulated? Was it efficient? How and why did it come to an end and with what consequences? The answers to these questions varied from one village to another across Europe and what follows is necessarily highly simplified (see de Moor et al. 2002, for a more detailed overview).

Common Land

The types of common land and the terminology used to describe such land varied across Europe.

Nevertheless, four major types of common land may be distinguished. First, the archetypical form of common land and the one with the widest geographical distribution is variously referred to as common waste, common pasture, waste, or common. This land was permanently common and most often grassland used for grazing animals. Usually such land was not suitable for arable cultivation typically because its natural fertility was low but sometimes for other reasons such a propensity to seasonal flooding. On some common wastes other resources were available, such as peat, turf, gorse or wood.

Second, in many parts of Europe much of the arable land (the land on which crops were grown) was also subject to common rights. Such land, known as open- fields, common fields, or common arable, was privately owned and cultivated but subject to common grazing. In its classic form each farmer held a number of long thin strips of land scattered over an extensive area and intermixed with the strips of other farmers. Each farmer cultivated his own crops on the arable. But when the harvest was over, or in years when the land was being fallowed, all those with common rights could turn their livestock into the fields to graze. Thus the open fields alternated between private and common land over the course of the agricultural cycle.

Third, common woodland for the production of fuel and timber was widespread on the European continent but unusual in England. This was similar to common waste in that it was permanently in common use.

Fourth, common meadows, which were permanent grasslands for the production of hay, were divided into separate blocks in private use but after the hay had been harvested were open to common grazing. Thus, like the open fields, common meadows alternated between private land and common land over the agricultural cycle.

Common Rights

As private property, the right to cultivate the common arable or to harvest the hay in common meadows lay with the owner of the land or the owner's tenant. Access to the common rights was considerably more complex and took different forms in different places; but it is possible to distinguish four main forms of access. First, in England and some parts of the Continent, the ownership or tenancy of particular buildings or landholdings was a prerequisite. Second, in many parts of the Continent citizenship of (as distinct from residence in) a commune or a municipality which itself owned the common resource was necessary, sometimes in combination with a property qualification. Third, in other parts of the Continent membership of a cooperative association which owned the common resources was necessary. Membership of these institutions was sometimes inherited, but sometimes it was attached to buildings or land (as in the first case). Fourth, there were cases were all residents in an area had common rights. But outside largely uninhabited areas, such as northern Sweden, this situation was unusual.

In consequence by no means all individuals or households enjoyed common rights. The proportion of the population that enjoyed common rights varied considerably from one region to another and changed over time. Where individuals or households did have common rights, the kinds and levels of the rights they enjoyed were determined by local regulations.

Regulation

Common land was almost invariably regulated by local institutions, often at the level of the individual village or manor. The institutions varied but were usually manor or village courts or village assemblies or committees of some kind, with the decisions made by a group of jurors. These institutions normally issued sets of rules, ordinances or by-laws which governed the usage of the commons and set fines for the infringement of rules. Officials or monitors were appointed to police the by-laws. The degree to which these institutions and their by-laws were subject to the influence of feudal overlords and the state varied considerably across Europe.

The by-laws provided the basic regulatory framework for managing the commons (for

examples of by-laws see Ault 1972). Their most critical function was to restrict the usage of common land and thus prevent a 'tragedy of the commons' developing. This was done in two ways. First, the by-laws would normally serve to restrict common rights to well-defined groups of users. For example, in much of England only those holding land in the open fields or with certain recognized dwellings, known as common-right houses, were allowed to pasture animals in the open fields or on the common pasture, while on much of the Continent pasture rights were restricted to citizens of communes or the holders of ancient farmsteads. Second, by-laws defined the amount of resources to which each commoner was entitled.

Thus, by-laws might specify the amount of peat or wood each commoner was entitled to dig or cut each year or the number and type of animals which could be kept, and for which months of the year they might be kept on the common pastures, open fields and common meadows.

The number of animals each commoner could put on the common land was generally controlled by one of two types of rules. One, known as 'stinting', simply specified the number and type of animals (the stint) which each commoner might keep on the common. Often the stint was proportional to the area or the value of land held. The other form of access, known as 'levancy' and 'couchancy', stated merely that each commoner could keep as many animals on the common as he could overwinter (that is, feed when the common was closed) on his own holding. How this was policed in practice is a moot point, but it certainly served to limit numbers and may have differed little from stinting in practice.

One consequence of these types of rule is that some individuals had no common rights at all. Another is that different individuals who did have common rights could have very different levels of access. The situation varied too much to allow generalization, beyond the suggestion that the level of inequality in England was probably greater and had proceeded further at an early date than anywhere else.

Enclosure

The process by which common land and common rights were abolished and replaced by recognizably modern forms of private property was part and parcel of a broader reform of landholding known as 'enclosure' which could also entail the consolidation of scattered holdings and the wholesale reallocation of land to create ring-fenced farms. Enclosure in some form is probably as old as common land itself. In England significant enclosure took place in the medieval period and from the 17th to the early 19th centuries. In most of Europe the widespread attack on common land began in the late 18th century in the wake of Physiocratic critiques. The later Napoleonic reforms and a subsequent series of state-sponsored drives to modernize agriculture in the 19th century led to more sustained enclosure. Some common land survives to this day, generally in mountainous areas.

Efficiency

By the 18th century common rights and common land were being widely criticized by agricultural improvers and others for restricting agricultural productivity. Most agricultural writers have accepted this view of common land as inefficient, and associated enclosure with major increases in productivity (Emle 1936; Chambers and Mingay 1966; Overton 1996). Common rights and common land imposed two kinds of limitation on agricultural improvement. First, the communal regulation of common land made it more difficult to introduce new agricultural techniques and technologies or to respond to changes in market opportunities. Second, the sharing of the outputs from common land made individual investment less attractive. The spread of nitrogen fixing crops and new drainage technologies, which often allowed the cultivation of formerly uncultivable common land, together with better transport links made enclosure a steadily more pressing issue in the 18th and 19th centuries.

A number of economic historians have reconsidered the inefficiency of open fields in an

English context. McCloskey (1976) has argued that the scattering of land in open fields in the medieval period was an efficient insurance against risk in a non-market economy. Allen (1992) has argued that enclosure did facilitate major technological changes obstructed by common land but that these innovations made only very marginal contributions to increased efficiency. Clark (1998) has argued that the inefficiencies imposed by common land were relatively modest and that, given the costs involved, enclosure was not economic until after 1750. However, the issue remains controversial essentially because it is inherently difficult to measure the agricultural productivity of farming in the 18th and 19th centuries with any degree of reliability. In other words, at present the data are too poor to allow an entirely plausible rebuttal of the views of 18th-century critics of the open fields. Moreover, much enclosure took place in the medieval period and in the 17th century (Wordie 1983) and any fully satisfactory theory of the efficiency or otherwise of open fields would need to be able to account for the longer-term chronology of enclosure. The persistence of open-field farming in France has been investigated by Grantham (1980) and Hoffman (1989).

Another controversial issue is the importance of common land to the poor. Many historians have argued that the poor derived considerable benefits from common land and that enclosure was socially damaging; but this remains controversial (see Neeson 1993; Shaw-Taylor 2001. The extent to which the poor benefited from common land and common rights is hard to reconstruct, poorly understood, and varied considerably across Europe.

Common-Pool Resources

This article has been concerned exclusively with common land and common rights as they existed in Europe before the 20th century. However, it should be noted that while open fields and common meadow may be peculiarly European forms, common waste and institutions for its management can be found all over the world. Analytically, these systems are part of a larger family of 'common-pool-resource' systems (Ostrom 1990) which have been adopted in many parts of the world to manage not just land but water resources and fish stocks as well.

C

See Also

- ▶ Access to Land and Development
- ▶ Common Property Resources
- ▶ Tragedy of the Commons

Bibliography

Allen, R. 1992. *Enclosure and the Yeoman: The agricultural development of the South Midlands, 1450–1850*. Oxford: Oxford University Press.

Ault, W. 1972. *Open-field-farming in medieval England: A study of village by-laws*. London: George Allen and Unwin.

Chambers, J., and G. Mingay. *1966. The agricultural revolution 1750–1880*. London: Batsford.

Clark, G. 1998. Commons sense: Common property rights, efficiency, and institutional change. *Journal of Economic History* 58: 73–102.

de Moor, M., L. Shaw-Taylor, and P. Warde, eds. 2002. *The management of common land in North West Europe, c. 1500–1850*. Turnhout: Brepols.

Emle, Lord. 1936. *English farming past and present*. 5th ed. London: Longmans, Green and Co..

Grantham, G. 1980. The persistence of open-field farming in nineteenth century France. *Journal of Economic History* 40: 515–531.

Hardin, G. 1968. The tragedy of the commons. *Science* 162: 1243–1248.

Hoffman, P. 1989. Institutions and agriculture in old-regime France. *Journal of Institutional and Theoretical Economics* 145: 166–181.

McCloskey, D. 1976. English open fields as behaviour towards risk. *Research in Economic History* 1: 124–170.

Neeson, J. 1993. *Commoners, common-right, enclosure and social change in England 1700–1820*. Cambridge: Cambridge University Press.

Ostrom, E. 1990. *Governing the commons: The evolution of institutions for collective action*. Cambridge: Cambridge University Press.

Overton, M. *1996. Agricultural revolution in England: The transformation of the agrarian economy 1500–1850*. Cambridge: Cambridge University Press.

Shaw-Taylor, L. 2001. Parliamentary enclosure and the emergence of an English agricultural proletariat. *Journal of Economic History* 61: 640–662.

Wordie, J. 1983. The chronology of English enclosure, 1500–1914. *Economic History Review* 36: 483–505.

Commons, John Rogers (1862–1945)

Warren J. Samuels

Keywords

American Association for Labor Legislation; American Economic Association; Capitalism; Collective action; Commons, J. R.; Fisher, I.; Institutional economics; Methodological individualism; National Bureau of Economic Research; New deal; Trade unions; Veblen, T

JEL Classifications

B31

Commons was born on 13 October 1862 in Hollandsburg, Ohio, and died on 11 May 1945 in Raleigh, North Carolina. He studied at Oberlin College (BA, 1888) and Johns Hopkins University (1888–90). He taught at Wesleyan, Oberlin, Indiana, Syracuse, and Wisconsin (1904–32).

The founder of the distinctive Wisconsin tradition of institutional economics, Commons derived his theoretical insights (generalized in his *Legal Foundations of Capitalism,* 1924, and *Institutional Economics,* 1934) from his practical, historical and empirical studies, particularly in the field of labour relations and in various areas of social reform. He drew insight not only from economics but also from the fields of political science, law, sociology and history. A principal adviser and architect of the Wisconsin progressive movement under Robert M. La Follette, Commons was active as an advisor to both state and federal governments. He was instrumental in drafting landmark legislation in the fields of industrial relations, civil service, public utility regulation, workmen's compensation and unemployment insurance. He served on federal and state industrial commissions, was a founder of the American Association for Labor Legislation, was active in the National Civic Federation, National Consumers' League (president, 1923–35), National Bureau of Economic Research (associate director, 1920–28), and the American Economic Association (president, 1917). He participated in antitrust litigation (especially the Pittsburgh Plus case) and in movements for reform of the monetary and banking system (often associated with Irving Fisher, who considered Commons one of the leading monetary economists of the period).

The critical thread uniting Commons's diverse writings was the development of institutions, especially within capitalism. He developed theories of the evolution of capitalism and of institutional change as a modifying force alleviating the major defects of capitalism. Commons came to recognize and stress that individual economic behaviour took place within institutions, which he defined as collective action in control, liberation, and expansion of individual action. The traditional methodologically individualist focus on individual buying and selling was not capable, in his view, of penetrating the forces, working rules and institutions governing the structural features of the economic system within which individuals operated. Crucial to the evolution and operation of the economic system was government, which was a principal means through which collective action and change were undertaken.

Commons rejected both classical harmonism and radical revolutionism in favour of a conflict and negotiational view of economic process. He accepted the reality of conflicting interests and sought realistic, evolutionary modes of their attenuation and resolution. These modes focused on a negotiational psychology in the context of a pluralist structure of power. He sought to enlist the open-minded and progressive leaders of business, labour and government in arrangements through which they could identify problems and design solutions acceptable to all parties.

In other contexts, he sought to use government as an agency for working out new arrangements to solve problems, such as worker insecurity and hardship, rather than promote systemic restructuring, although to many conservatives his ventures were radical enough. To these ends Commons and small armies of associates engaged in fact finding – his look-and-see methodology – in a spirit of bringing all scientific knowledge to bear

on problem solving. From these experiences, indeed already manifest in the underlying strategy, Commons developed a theory of government as alternately a mediator of conflicting interests and an arena in which conflicting interests bargained over their differences; a theory of the complex organization – in terms of freedom, power and coercion – and evolution of the legal foundations of capitalism, which centred in part on the composing of major structural conflicts through the mutual accommodation of interests; and a theory of institutions with an affirmative view of their roles in organizing individual activity and resolving conflict.

The institutions Commons studied most closely were trade unions and government, particularly the judiciary. He developed his theory of the economic role of government in part on the basis of his study of the efforts of workers to improve their market position and in part on the use of government by both enemies and friends of labour. Commons's was an interpretation of trade unions as a non-revolutionary development, as collective action seeking to do for workers what the organizations of business attempted to do for their owners and managers. His study of the reception given unions and reform legislation led him to recognize the critical role of the United States Supreme Court (and the courts generally), and its conception of what was reasonable in the development and application of the working rules which governed the acquisition and use of power in the market. Accordingly, Commons developed a theory of property which stressed its evolution and role in governing the structure of participation and relative withholding capacity in the market.

Commons also developed a theory of institutions which focused on their respective different mixtures of bargaining, rationing and managerial transactions, all taking place within a legal framework which was itself subject to change.

Although Commons's institutionalism had different emphases from that of Thorstein Veblen, for example, in that Commons stressed reform of the capitalist framework, they shared a view of economics as political economy and of the economy as comprising more than the market. Unlike Veblen, Commons was not antagonistic toward businessmen, and indeed accepted capitalism, though not necessarily on the terms given or preferred by the established power structure.

Commons was one of the few American economists to found a 'school', a tradition that was carried forward by a corps of students, especially Selig Perlman, Edwin E. Witte, Martin Glaeser and Kenneth Parsons. Much mid-20th-century American social reform, the New Deal for example, drew on or reflected the work of Commons and his fellow workers and students.

Selected Works

1893. *The distribution of wealth.* New York: Macmillan.
1905. *Trade unionism and labor problems.* Boston: Ginn.
1910–11. (With others.) *Documentary history of American industrial society.* 10 vols. Cleveland: A.H. Clark.
1916. (With J.B. Andrews.) *Principles of labor legislation.* New York: Harper.
1919. *Industrial goodwill*. New York: McGraw-Hill.
1919–1935. (With others.) *History of labor in the United States.* 4 vols. New York: Macmillan.
1921. *Industrial government.* New York: Macmillan.
1924. *The legal foundations of capitalism.* New York: Macmillan.
1934. *Institutional economics.* New York: Macmillan.
1934. *Myself.* New York: Macmillan.
1950. *The economics of collective action.* Madison: University of Wisconsin Press.

Communications

Roger G. Noll

The economics of communications is a loose, somewhat vaguely defined amalgam of topics in applied microeconomics. Although having close

ties to the microeconomic theory of the economics of information, it is probably best characterized as a subfield of industrial organization, regulation and public enterprise that deals with the communications sector: telecommunications, broadcasting, the print media, the performing arts and the postal system. Of course, the activities that constitute this list are somewhat arbitrary, but they reflect what is both taught and studied by people in the subfield as well as some important economic realities that make specialized studies of the communications sector a valid category among distinct intellectual pursuits. First among these realities is that the industries in the communications sector are closely linked. Broadcasting competes with the performing arts for both audience and inputs, and telecommunications competes with the postal service. Moreover, telecommunications networks are capable of delivering broadcast services, and vice versa. Among the products over which the postal system, telecommunications and cable television compete is the delivery of the output of the print media.

Another unifying theme across communications industries is the connection of the study of the sector to the economics of public goods and externalities. Communications is the production and dissemination of information. Some aspects of the production of information are public goods, and the dissemination and use of information can have important external effects. Moreover, most of these external effects are non-economic phenomena such as political participation, the cultural values held by members of a society or the level of violence. Because of the unique character of these externalities, the motives for public policy in communications are closely linked with a society's fundamental political and social values. Thus, freedom of speech and the extent of the right to privacy, as well as the use of control of communications to manipulate the political process, are at the heart of debates over communications policy.

Rationales for Government Intervention

Not surprisingly, the role of the public sector is very large in the communications sector in nearly every nation. Subsidization, nationalization and extremely detailed regulation of prices and attributes of the product are common. In market-oriented societies, telecommunications and mail are nationalized or subject to economic regulation for much the same reasons that underpin the same policies in other infrastructural industries: that these industries are natural monopolies and that their performance and pattern of development profoundly affect the development of much of the rest of the economy. But even here, unique externality arguments are brought forth as additional factors to be taken into account by policymakers. First, subscription to the telecommunications network or access to mail delivery creates the capability to receive communication from others. A person who decides to mail a letter or place a telephone call presumably considers only his or her net benefits from the communication; the willingness to pay of the recipient (positive or negative) is not taken into account. Thus, for example, the extent of phone service and the pattern of calling can be expected to be inefficient if each person bears the full cost of, first, subscribing to the network, and then placing telephone calls. In particular, if some potential customers have too low a willingness to pay to become subscribers, but are also desired objects of communications by others, the number of subscribers will be too low if subscribers must bear the full cost of connecting them to the network. This argument constitutes the foundation for the 'universal service' objective; that is, a policy of maximizing the number of subscribers to the telephone network, and the policy practised nearly everywhere of adopting a price structure for telephone services that subsidizes installation charges, pay-telephone prices and/or monthly access charges to the local network, especially for customers in high-cost areas such as rural communities.

A second externality of the telecommunications system is said to be its contribution to national security. A joint product of a private telecommunications network is a ready resource that can be commandeered and used by government in times of national emergency, such as foreign attack, natural disasters or accidents. The use of communications to coordinate a response to

such an event, then, should play a role in affecting the capacity and design of the telecommunications system, and often is the basis of an argument for building into the system more redundancy and interconnectability than might otherwise be optimal and than independent private concerns would undertake on their own. These contingent needs by government have been said to constitute a separate natural-monopoly argument, an example of 'economies of scope' between private and public uses that can only be captured if the private system is a single, integrated whole. In the United States, for example, the Department of Defense was a consistent critic of proposals to relax regulation and increase competition in the telecommunications industry during the 1960s and 1970s.

The externalities associated with the mass media have to do with the social, political and psychological consequences of the content of information, and for the most part are dealt with by scholars from disciplines other than economics. (The main exception is research on the effects of advertising, where an inconclusive debate has raged for decades as to whether the informational value of advertising exceeds the sum of its direct costs and possible resource misallocation owing to manipulation and misperceptions of consumers.) The analytical foundation for the belief in the importance of informational externalities is the proposition that people's behaviour as citizens, parents, consumers, workers, friends, and so on, can be significantly affected, at least in the short run, by the informational content of the mass media. Once one accepts this proposition, the next logical step is to entertain the idea that censorship by the state, at least in principle, can prevent some of the external diseconomies of destructive content, while proactive state interventions to channel the content of the media towards greater educational and otherwise uplifting content can provide additional social benefits.

The most obvious manifestations of these ideas are in broadcasting. The British Broadcasting Corporation was founded on openly paternalistic principles about the potential of radio broadcasts for educational and other uplifting purposes. Until the recent move towards decentralization through cable television and towards private, commercial television, a core principle of French broadcasting policy was to preserve French culture and values by limiting and censoring programmes from other nations. In Germany, decentralized, regional quasi-public broadcast monopolies were created after World War II to protect simultaneously against capture by the national government or by the national print media barons, either of which, it was feared, might use broadcasting to arouse nationally destructive political passions. And in the United States, broadcast licensing has, until recently, enforced a long set of standards for evaluating competitors for a given license, including personal characteristics and other business holdings of the licensee and both performance and promise about the extent of 'public service' programming offered by commercial as well as non-commercial (educational) outlets.

Of course, other mass media are not free of similar policy constraints, although the print media and the arts are usually accorded greater freedom in the content of their messages than are broadcasters. The areas of policy controversy are the definition of the liability for slanderous attacks and the concomitant definition of privacy rights, the boundaries between pornography and legitimate expression, and the principles separating sedition from reasonable political discourse.

The core economic issue in this debate is whether the 'marketplace for ideas' works well without intervention, or at least better than is the case with active political intervention by the very government authorities whose security and power can be affected by the content of communications. The argument for non-intervention is twofold. Positively, it is that in the end people's tastes in ideas should be accorded the same status as their tastes in other goods as long as the consumption of communications produces no external diseconomies. If communications cause bad behaviours, then if people are informed about this fact – and about the punishments exacted if those behaviours are manifest – they will efficiently anticipate this in making decisions about which communications to receive and how to treat those that are received. And, as with goods, ideas about how the world works that prove correct will be perceived, at least eventually, as superior to less correct ideas.

Negatively, the argument for an unregulated marketplace for ideas is a pessimistic forecast of how political intervention is likely to work: a combination of orientation towards propaganda to serve the interests of preserving the status quo and an extreme sensitivity to either vocal, organized single-issue groups seeking to impose their values on others or a tyranny of the majority that persecutes those who stray too far from current norms.

The other side of the dispute, usually advocated more by non-economists, is rooted in observed relationships between communications and behaviour, perhaps best documented in the study of the effects of television on violence (especially by children) and on the manipulation of the news for short-term political objectives. This position regards the efficiency of the marketplace for ideas as demonstrably poor, at least in the short run; implicitly, it accords less credence to the proposition that individuals are as rational – indeed, are even proactive – in selecting among competing communications as economic theory assumes. Proponents of intervention especially emphasize the unformed and manipulable attitudes of children.

The final pervasive feature in the communications sector that deserves further elaboration is the partially non-rivalrous nature of the consumption of information. All information is a public good in the sense that once a new information product has been created for a first user, it does not have to be created again for subsequent users: in principle, at least, the first use of information does not preclude its use by others. In practice, this characteristic may be unimportant. Information must be disseminated in some way to subsequent users, and the cost of dissemination may exceed the cost of secondary creation – as for example can be the case for a simple computer program. Or, information may be very cheaply privatized so that the public goods characteristic introduces no significant inefficiency to a private market system of distribution. Nevertheless, the publicness of information is a serious issue in an assessment of the performance of allocational institutions in the communications sector, and in the design of private market processes for allocating resources the problems of publicness must be addressed. Whether the product is a written news report, a novel, a theatrical production, a television broadcast, or 'Dial-A-Joke' on the telephone, the problem is fundamentally the same: producers will not supply a product unless they can recover the opportunity cost for creating it, yet the marginal cost of providing the product to one more consumer does not include any of the production costs of the information. Hence, efficient provision of information requires one of the following: subsidies of the production of information, or price discrimination with protection against arbitrage so that consumers with relatively low willingness to pay for information will not be inefficiently excluded. In practice, both are common. Governments subsidize broadcasting and performing arts by direct payments, and certain users of the postal and telecommunications system either directly or by engaging in price discrimination (e.g. the differences in basic monthly service rates of telephones between residences and businesses, and the lower postal rates for circulating the print media). Of course, neither direct subsidies nor discriminatory prices are explicitly designed in a quantitative sense to offset the inefficiencies of private provision of public goods, so that the issue of optimal pricing of communications services is an active and still-developing field of research. The focus here is on the two fields in which most of the work has been done: telecommunications and broadcasting. Moreover, the discussion includes research on market structure issues because of their close connection to the implications of alternative pricing policies.

Pricing and Market Structure in Telecommunications

The telecommunications industry in the United States offers an array of services that until very recently were provided as joint products by a legally protected monopoly. When the monopoly was secure and unquestioned, the pricing problem was to devise a price structure that recovered joint and fixed costs with minimal loss of efficiency. As the natural monopoly presumption came to be

called into question, the pricing problem began to incorporate another dimension, to provide appropriate signals to potential competitors so that the market structure would evolve efficiently.

To understand the rudiments of the telephone pricing problem requires some basic knowledge about the technical characteristics of the telecommunications network. The traditional telephone system is best conceptualized as having four components: customer terminal equipment (a telephone, a computer terminal, a switchboard); a pair of copper wires connecting each terminal device to a central switch; the central switch that serves the local community; and a hierarchy of transmission conduits and additional switches that serve to connect the local switches. Typically the telephone price structure has three elements: an installation charge for activating a customer's copper wire pairs; a basic monthly service charge for renting terminal equipment and the copper wire pairs; and a message toll for placing telephone calls. The common practice is for the installation charge to cover only a fraction of installation costs as a means of encouraging universal service, and for the basic monthly charge to entitle the subscriber to unlimited local calling – sometimes not confined to other telephones connected to the same local switch, but also including calls through adjacent local switches. Usually the basic monthly charge is much higher (by a factor of 2 or 3) for business than for residences, but within each of these categories it tends to be approximately the same over wide geographic areas regardless of differences in cost of service.

Until about 1960, the revenues from installation charges and the basic monthly rate approximately covered the cost of local service (including local switches). But as long-distance toll calls became more important, telephone companies increasingly used toll revenues to cover part of the cost of the local system. This required no increase in toll prices; indeed, long-distance prices generally were falling because technological change was extraordinarily rapid in this segment of the system. By simply letting prices fall a little more slowly than costs, a large and growing fraction of local network costs could be paid for by toll. These revenues could be used to construct systems in high-cost rural areas without causing increased prices for basic service elsewhere, again to encourage universal service.

Since toll calls pass through local switches they impose a cost on the local network, because local switches must be designed to be large enough and complex enough to accommodate them. Consequently, toll prices would bear some local system costs in an efficient pricing structure. In addition, however, toll calls also contributed to 'non-traffic sensitive' (NTS) costs – the terminals and copper wire – even though, by definition, the magnitude of investment required for this equipment was unrelated to the amount of calling.

Obviously, this pricing structure not only encouraged universal service but encouraged local calling (with a zero price at the margin) while discouraging long-distance calling compared to an efficiency standard. Encouraging subscriptions to the system may be warranted on efficiency grounds, although the magnitude of the subscription externality has not been quantified, and so it is not possible to tell whether the amount of the subsidy is justified. Likewise, a subsidy of local calls may be desirable, but the method of subsidization is of doubtful validity. The external benefit (or cost) of a call falls on the person being called, not on society generally.

Hence, the optimal pricing structure involves a sharing of the costs of calling between the parties to a conversation, where the costs involve the operating costs of the system and the effect of calls on the required capacity of the switching system. Only if metering costs were large in comparison with the costs of calling would it make sense not to charge for calls, but with modern electronic switching metering costs are not significant, so that one cannot justify a subsidy for local calls. Moreover, even if one could, there is no justification for taxing long-distance calls to pay the subsidy unless one believes that the externality of a local call is substantially more important than the externality of a long-distance call.

The general structure of an optimal price structure for the telephone network, given important externalities and natural monopoly, can be derived as follows. Begin with a basic monthly charge that

would pay the marginal cost of terminals and copper wire connections to the local switch, and toll charges on all calls that pay the marginal cost of the switching and transmission facilities that are traffic sensitive. These prices need further adjustment, for they may collect too much or too little total revenue. But prior to this adjustment they must also be uniformly adjusted downwards to account for the externality of subscribing and calling (assuming that people like to receive phone calls). The adjustment for the toll rate can simply be passed on to the recipient of the call; however, the basic monthly rate must come down for everyone. At this point the likely case is that the basic monthly rate does not cover the NTS costs, so that further adjustments must be made. One possibility is a subsidy paid from an economy-wide tax, but more likely the additional revenues will come from the rate structure of the telephone company. The first-best solution is to raise infra-marginal prices, such as the cost of the first few calls made per month, producing what amounts to quantity discounts for all types of calls. Alternatively, one could adopt Ramsey pricing, raising the price the most for services with relatively inelastic demand.

The resulting price structure would have a number of very interesting features. Call recipients would pay a share of the price of a call. To implement this so as only to charge for calls with a positive benefit, the shared cost would start a decent interval after the call is answered, and subscribers who desired it could be permitted to designate in advance that they would bear the full cost of their calls. All prices would be built on marginal costs, which means peak-load pricing of calling and prices for both basic access and calling that were higher in high-cost areas. To the extent that Ramsey prices were invoked, they would most certainly rely primarily on basic monthly charges, for this has by far the lowest demand elasticity: estimates range between -0.02 and -0.10. Thus, even if there is a significant externality associated with subscribing to the network, the Ramsey pricing method for paying for it involves raising the price of basic access. Or, putting the matter another way, ignoring this externality in setting prices will have very little effect on subscriptions to the system, and hence very little effect on efficiency. Finally, differences between residential and business basic monthly charges would exist only if their externality value differed, they imposed different costs on the system or they had different price elasticities.

Obviously, the pricing structure of telephone service has never reflected these principles. Until the 1970s, government officials perceived the extent of inefficiency of the pricing structure as something of an academic issue and largely ignored it. But technological change and the false signals to entrants from the price structure have led to strong pressures for competitive entry into the formerly secure telephone monopoly. Computer technology has vastly diversified the demand for telecommunications, as well as vastly increased its magnitude, and computers and other advances in microelectronics have altered the technology of supply. Examples of the broad range of new computer-based services include on-line connections to mainframes and data bases for technical and business use, automatic teller machines, remote sensing for protection against fires and burglars, and reservation services. Each of these uses has somewhat different technical requirements, so that the optimal market structure for the industry may well be to have a product-differentiated oligopoly, even if each has unexploited economies of scale. Moreover, the greater demand created by these technical advances allows considerable exploitation of scale economies even in a segmented system.

On the supply side, advances in electronics have changed the basic character of the local network. No longer are high-density networks built of dedicated copper wire pairs for each terminal. Instead, micro-electronic technology allows multiple signals on the same wires, and small-scale switches distributed throughout the local network that serve to concentrate lines from many terminals into a smaller number of active circuits, taking advantage of the fact that not all terminals are in use simultaneously. This reduces the unit cost of capacity and hence the significance of scale economies in the local network. Moreover, it undermines a cornerstone of the optimal-pricing structure that was developed

above by eliminating most of the NTS costs. If, as is becoming the case, customers own their terminal equipment, and if line concentration begins when a small number of terminals are aggregated into a single pathway to the first switch, then nearly all of the system that is owned by the local telephone company consists of traffic-sensitive investment. Hence, the trend should be away from reliance on the basic monthly charge and towards greater reliance on message tolls for calling.

The failure to adjust the pricing system to the realities of costs and technology adds to the pressure for competitive entry in the parts of the system where prices are higher than the costs of service. Specifically, the attempt to tax long-distance service in order to subsidize local calling makes relatively intense users of long-distance service ripe candidates for a competitive long-distance supplier. Large companies with many telephone lines who can provide their own concentrated connections between their facilities have a strong incentive to bypass subscriptions to the local network. And other electronic pathways for communications, such as cable television and point-to-point uses of the radio spectrum, can be exploited to bypass the telecommunications network.

Thus far, government officials responsible for telecommunications policy whether as operators of public enterprises or regulators of private utilities, have focused more on the structural aspects of the problem than on pricing issues. Even in the United States, which has perhaps the greatest commitment to competition, government policy regarding entrants has been the binding constraint on the growth of competition, and the price structure is still replete with inefficiencies. The likely explanation for this phenomenon is the belief by political actors that the cost of local service to residences is the price that is most visible politically and that rationalized pricing will cause residential service to become more expensive, either from raising basic monthly charges or from message toll for local calls. To avoid raising residential price, political actors therefore believe that they have to keep some other prices above the cost of service, and to maintain these prices in the face of diminished or perhaps even non-existent natural monopoly they must erect barriers to competitive entry aimed at the overcharged customers.

Prices and Market Structure in Broadcasting

The most common way to pay for broadcasting is to provide signals to the audience at no charge, and to rely on either advertising or government subsidies as to the source of revenues. In one sense such an arrangement seems to fit the fact that broadcasts are a classic public good; the marginal cost to the broadcaster of one more person receiving a broadcast is zero, and consumption among members of the audience is completely non-rivalrous. Hence, any attempt to charge a viewer for a programme can introduce inefficiency to the extent that anyone is thereby excluded from participation who also has a positive willingness to pay to join the audience.

The difficulty with free broadcasting, however, is that it does not necessarily result in programmes that maximize the net willingness to pay of the audience. Ignoring for a moment the frictions in the political process and the incentives of political actors to manipulate programme content to their private benefit, both subsidized and advertiser-supported television lead broadcasters to measure their success primarily on the basis of the size of audience. In a subsidized system, the objective would be to make certain that political support is as high as possible, and in an advertising system, in which the broadcaster is selling the attention of viewers, revenues are more or less proportional to audience size. The issue in both cases is not whether audience satisfaction is maximized but whether it is kept high enough for a large enough number of people to maximize revenues from a payment system that is not based on the intensity of preferences but on the number of satisfied customers. In particular, small groups with intense willingness to pay for an unusual type of programming material will generally not have their preferences satisfied even if their aggregate

willingness to pay exceeds that of a large audience for the traditional mass-audience programme.

Three means are available for coping with this state of affairs. One is to expand the number of broadcast options until all groups are satisfied. Suppose that there is a large mass audience and a series of small, specialized ones. As the number of stations expands, the audience for mass programmes each can expect will be the total mass audience divided by the number of stations. Eventually, there will be enough stations so that the largest specialized taste will constitute a larger audience than the share of the mass audience the next station could expect to capture, so that a strategy to maximize audience will lead to specialization. In the United States, this is more or less the policy with respect to radio broadcasting. In the early years of radio, the Federal Communications Commission tried to assure diversity in commercial broadcasting by specifying the format (e.g. type of programmes) that a station could broadcast. Recently, station formats have been deregulated, yet the multiplicity of categories remain, in much the same fashion that there is a broad spectrum of magazines and books by type of material.

In television, the strategy of increasing the number of stations is more difficult to follow. Television stations consume far more radio frequency space than radio stations, and no nation has thus far been willing to allocate enough high-quality radio spectrum to television to provide much of a test as to whether specialization might take place in a more extensive industry. An unplanned test, however, is under way in Italy, where in the 1970s the courts declared that the government had no constitutional right to limit the number of television stations, and largely unregulated entry has taken place on a massive scale. It is too soon to tell what the ultimate outcome of this system will be.

The second mechanism to produce more diversity in television is to allow the audience to express a willingness to pay, commonly by installing cable television with much higher capacity than off-air television and charging customers on the basis of the number and type of channels that they elect to receive. The inefficiency inherent in this system is the cost of privatizing broadcasts so that on either a per programme or per channel basis they can be sold. Prior to the extensive development of cable television in the United States, the common speculation was that privatization of broadcasts would cause a diversion from traditional mass audience programming, with more activity in cultural programmes, educational broadcasting and public affairs. The expectation was based upon the belief that higher-income groups were more interested in diversity and would have more influence in determining the content of for-pay systems. In practice, this expectation has not been realized. The new cable-oriented networks for the most part offer programming that is like that provided by off-air broadcasters, such as movies, sports events and regular series. The principal exception is in public affairs, where national cable news and public events networks have succeeded. Educational and cultural programming, however, has been largely unsuccessful. The inference to be drawn is that scarcity in television stations caused an excess demand for television, but primarily for programmes that were much like those featured by off-air stations, largely oriented towards the mass audience.

The third means for increasing diversity is to create a single, multi-outlet monopoly broadcasting entity. If such an entity seeks to maximize total audience, it will not have as much incentive to duplicate mass programming, because audience substitution from one channel to the next will have no value. A second or third mass-audience station will increase the size of the total audience, but the evidence indicates that the effect is small compared to audience diversion. For example, in the United States the first television station captures a little less than half of the potential audience available in prime evening viewing time. No matter how many additional stations are added, the maximum viewing share appears to be about 80%, and this is almost totally achieved after three or four stations are operating. This suggests that a multichannel monopolist would either diversify programming on the second or third channel, or simply elect not to broadcast on more than one or two channels, depending on the relationship

between the value of a net increment to the audience and the costs of adding another channel.

American public television provides an example of a novel method of support, for it is one of the few attempts to implement a decentralized decision process for acquiring a public good (here programmes). The first component of the system is the method of public financing, which involves multiple year, advance funding to erect some barrier to political manipulation of programmes. The public funds are then divided into three components: a budget for experimental programming that is spent by an independent, quasi-public entity (the Corporation for Public Broadcasting); a budget for the technical operation of the national network (Public Broadcasting Service); and a direct subsidy of local stations. This subsidy is based in part on the success of the station in obtaining private contributions from its audience. Thus the station subsidy amounts to an attempt to overcome the free-rider problem faced by viewers by providing matching funds for their contributions.

The second component of the system is the mechanism by which stations decide which programmes will appear on the network. This is accomplished by a combination voting and price system. The price of a programme for each station is determined by the size of the community it serves, and stations then vote on each programme proposal. If some stations vote against the proposal, the prices faced by the supporters are increased by an amount necessary to allow the programme to cover its costs, and voting proceeds again. The process continues until programmes are either purchased or discarded; usually fewer than a dozen iterations are required to reach a decision. Stations voting against a programme are excluded from broadcasting it; however, stations may later join the group paying for it by paying a premium price.

The programme acquisition process decentralizes network programming to the stations, thereby serving two ends. First, because the station budgets depend on contributions, or the voluntary willingness to pay of the audience, there exists a feedback mechanism from the audience to the network that is similar to a pay-TV system. Second, the network schedule becomes less vulnerable to political attack, for centralized government officials who might seek to control it face a collective of over 150 station licensees, who, in turn, are actively using contributions patterns to make decisions about which programmes to acquire.

The American system of financing public television does not, of course, have pristine efficiency properties; neither the voluntary audience contributions nor the mechanism whereby stations select programmes is an incentive-compatible mechanism. Nevertheless, in the inherently imperfect world of public goods acquisition, they appear to perform remarkably well, and experimental investigations in a laboratory setting suggest that the method of acquiring programmes can be productively employed in a variety of settings for collective decisions.

Remaining Issues

Two aspects of the communications sector make it a ripe area for continuing study. First is the rapidly evolving technology of supply and demand, and the second is the pervasive and changing influence of political processes on the structure and performance of the sector.

With changing technology has come a significant change in the pattern of demand for services. This suggests that historical patterns of use and estimates of service-specific demand are unreliable predictors of the future. Yet relatively little research has investigated how changing technology – lower costs, greater possibilities of use, more technical capabilities – have affected key aspects of demand: the rate of growth by service and customer category and the own-and cross-elasticities of demand.

Changing technical possibilities and demand should also feed back into the political forces that guide the development of the sector. Most advanced industrialized nations are in the midst of transition in at least some policies regarding communications, such as the privatization and introduction of competition in telecommunications in Japan and Great Britain, and the

elimination of the state broadcasting monopolies in France and Italy. These changes deserve study on two counts: how these dramatic policy changes affect performance, and what political forces they may be creating that will shape policy and industry structure in the future.

A period of rapid change is one in which important new knowledge is likely to be forthcoming. One can anticipate that a summary of the economics of communications a decade or two hence will contain significant and surprising new insights.

See Also

► Public Utility Pricing

References

Bloch, H., and M. Wirth. 1984. The demand for pay services on cable television. *Information Economics and Policy* 1(4): 311–332.

Brock, G. 1981. *The telecommunications industry: The dynamics of market structure*. Cambridge, MA: Harvard University Press.

Coase, R. 1959. The Federal Communications Commission. *Journal of Law and Economics* 2(1): 1–40.

Courville, L., A. de Fontenay, and R. Dobell. 1983. *Economic analysis of telecommunications*. Amsterdam: North-Holland.

Evans, D. (ed.). 1971. *Breaking up bell*. Amsterdam: North-Holland.

Levin, H. 1971. *The invisible resource*. Baltimore: Johns Hopkins Press.

Machlup, F. 1980. *The production and distribution of knowledge in the United States*. Princeton: Princeton University Press.

Mitchell, B. 1978. Optimal pricing and local telephone services. *American Economic Review* 68(4): 517–537.

Network Inquiry Special Staff. 1980. *New television networks: Entry, jurisdiction, ownership and regulation*. Washington, DC: Federal Communications Commission.

Noll, R. 1985. 'Let them make toll calls': A state regulator's lament. *American Economic Review* 75(2): 52–56.

Noll, R., M.J. Peck, and J.J. McGowan. 1973. *Economic aspects of television regulation*. Washington, DC: Brookings.

Owen, B. 1975. *Economics and freedom of expression*. Cambridge, MA: Ballinger.

Owen, B., J. Beebe, and W. Manning. 1974. *Television economics*. Lexington: D.C. Heath.

Park, R.E. 1972. Prospects for cable in the 100 largest television markets. *Bell Journal of Economics* 3(1): 130–150.

Park, R.E. 1975. New television networks. *Bell Journal of Economics* 6(2): 607–620.

Rosse, J. 1967. Daily newspapers, monopolistic competition, and economies of scale. *American Economic Review* 52(2): 522–533.

Rosse, J., J. Dertouzos, M. Robinson, and S. Wildman. 1979. Economic issues in mass communications industries. In *Proceedings of the symposium of media concentration*, vol. I, 40–192. Washington, DC: Federal Trade Commission.

Snow, M. 1986. *Marketplace for telecommunications: Regulation and deregulation in industrialized democracies*. White Plains: Longman.

Spence, A.M., and B. Owen. 1977. Television programming, monopolistic competition and welfare. *Quarterly Journal of Economics* 91(1): 103–121.

Spitzer, M. 1985. Controlling the content of print and broadcast. *Southern California Law Review* 58(6): 1349–1405.

Steiner, P. 1952. Program patterns and preferences, and the workability of competition in radio broadcasting. *Quarterly Journal of Economics* 66(2): 194–223.

Taylor, L. 1980. *Telecommunications demand: A survey and critique*. Cambridge, MA: Ballinger.

von Weiszacker, C. 1984. Free entry into telecommunications? *Information Economics and Policy* 1(3): 197–216.

Communism

Ernest Mandel

The term 'communism' was first used in modern times to designate a specific economic doctrine (or regime), and a political creed intending to introduce such a regime, by the French lawyer Etiénne Cabet in the late 1830s; his works, especially the utopia *L'Icarie*, were influential among the Paris working class before the revolution of 1848. In 1840, the first 'communist banquet' was held in Paris – banquets and banquet speeches were a common form of political protest under the July monarchy. The term spread rapidly, so that Karl Marx could entitle one of his first political articles of 16 October 1842 'Der Kommunismus und die Augsburger *Allgemeine*

Zeitung'. He noted that 'communism' was already an international movement, manifesting itself in Britain and Germany besides France, and traced its origin to Plato. He could have mentioned ancient Jewish sects and early Christian monasteries too.

In fact, some of the so-called 'Utopian socialists', in the first place the German Weitling, called themselves communists and spread the influence of the new doctrine among German itinerant handicraftsmen all over Europe, as well as among the more settled industrial workers of the Rhineland. Under the influence of Marx and Engels, the League of the Just (Bund dez Gerechten) they had created, changed its name to Communist League in 1846. The League requested the two young German authors to draft a declaration of principle for their organization. This declaration would appear in February 1848 under the title *Communist Manifesto*, which would make the words 'communism' and 'communists' famous the world over.

Communism, from then on, would designate both a classless society without property, without ownership – either private or nationalized – of the means of production, without commodity production, money or a state apparatus separate and apart from the members of the community, and the social-political movement to arrive at that society. After the victory of the Russian October revolution in 1917, that movement would tend to be identified by and large with Communist parties and a Communist International (or at least an 'international communist movement'), though there exists a tiny minority of communists, inspired by the Dutch astronomer Pannekoek, who are hostile to a party organization of any kind (the so-called 'council communists', *Rätekommunisten*).

The first attempts to arrive at a communist society (leaving aside early, medieval and more modern christian communities) were made in the United States in the 19th century, through the establishment of small agrarian settlements based upon collective property, communally organized labour and the total absence of money inside their boundaries. From that point of view, they differed radically from the production cooperatives promoted for example by the English industrialist and philanthropist Robert Owen. Weitling himself created such a community, significantly called Communia. Although they were generally established by a selected group of followers who shared common convictions and interests, these agrarian communities did not survive long in a hostile environment. The nearest contemporary extension of these early communist settlements are the *kibbutzim* in Israel.

Rather rapidly, and certainly after the appearance of the *Communist Manifesto*, communism came to be associated less with small communities set up by morally or intellectually selected elites, but with the general movement of emancipation of the modern working class, if not in its totality at least in its majority, encompassing furthermore the main countries (wealth-wise and population-wise) of the world. In the major theoretical treatise of their younger years, *The German Ideology*, Marx and Engels stated emphatically:

> Empirically, communism is only possible as the act of dominant peoples 'all at once' and similtaneously, which presupposes the universal development of productive forces and the world intercourse bound up with them. . . .The proletariat can thus only exist worldhistorically, just as communism, its activity, can only have a 'world-historical' existence.

And, earlier in the same passage,

> . . . This development of productive forces (which at the same time implies the actual empirical existence of men in their world-historical, instead of local, being) is an absolutely necessary practical premise, because without it privation, is merely made general, and with want the struggle for necessities would begin again, and all the old filthy business would necessarily be restored . . . ([1845–6] 1976, p.49).

That line of argument is to-day repeated by most orthodox marxists (communists), who find in it an explanation of what 'went wrong' in Soviet Russia, once it was isolated in a capitalist environment as a result of the defeat of revolution in other European countries in the 1918–1923 period. But many 'official' Communist Parties still stick to Stalin's particular version of communism, according to which it is possible to

successfully complete the building of socialism and communism in a single country, or in a small number of countries.

The radical and international definition of a communist society given by Marx and Engels inevitably leads to the perspective of a *transition* (transition period) between capitalism and communism. Marx and Engels first, notably in their writings about the Paris Commune – *The Civil War in France* – and in their *Critique of the Gotha Programme* [of the German social-democratic party], Lenin later – especially in his book *State and Revolution* – tried to give at least a general sketch of what that transition would be like. It centres around the following ideas:

The proletariat, as the only social class radically opposed to private ownership of the means of production, and likewise as the only class which has potentially the power to paralyse and overthrow bourgeois society, as well as the inclination to collective cooperation and solidarity which are the motive forces of the building of communism, conquers political (state) power. It uses that power ('the dictatorship of the proletariat') to make more and more 'despotic inroads' into the realm of private property and private production, substituting for them collectively and consciously (planned) organized output, increasingly turned towards direct satisfaction of needs. This implies a gradual withering away of market economy.

The dictatorship of the proletariat, however, being the instrument of the majority to hold down a minority, does not need a heavy apparatus of full-time functionaries, and certainly no heavy apparatus of repression. It is a state *sui generis*, a state which starts to wither away from its inception, i.e. it starts to devolve more and more of the traditional state functions to self-administrating bodies of citizens, to society in its totality. This withering away of the state goes hand in hand with the indicated withering away of commodity production and of money, accompanying a general withering away of social classes and social stratification, i.e. of the division of society between administrators and administrated, between 'bosses' and 'bossed over' people.

That vision of transition towards communism as an essentially evolutionary process obviously has preconditions: that the countries engaged on that road already enjoy a relatively high level of development (industrialization, modernization, material wealth, stock of infrastructure, level of skill and culture of the people, etc.), created by capitalism itself; that the building of the new society is supported by the majority of the population (i.e. that the wage-earners already represent the great majority of the producers and that they have passed the threshhold of a necessary level of socialist political class consciousness); that the process encompasses the major countries of the world.

Marx, Engels, Lenin and their main disciples and co-thinkers like Rosa Luxemburg, Trotsky, Gramsci, Otto Bauer, Rudolf Hilferding, Bukharin et al. – incidentally also Stalin until 1928 – distinguished successive stages of the communist society: the lower stage, generally called 'socialism', in which there would be neither commodity production nor classes, but in which the individual's access to the consumption fund would still be strictly measured by his quantitative labour input, evaluated in hours of labour; and a higher stage, generally called 'communism', in which the principle of *satisfaction of needs* for everyone would apply, independently of any exact measurement of work performed. Marx established that basic difference between the two stages of communism in his *Critique of the Gotha Programme*, together with so much else. It was elaborated at length in Lenin's *State and Revolution*.

In the light of these principles, it is clear that no socialist or communist society exists anywhere in the world today. It is only possible to speak about 'really existing socialism' at present, if one introduces a new, 'reductionist' definition of a socialist society, as being only identical with predominantly nationalized property of the means of production and central economic planning. This is obviously different from the definition of socialism in the classical marxist scriptures. Whether such a new definition is legitimate or not in the light of historical experience is a matter of political and philosophical judgement. It is in any case another matter altogether than ascertaining

whether the radical emancipatory goals projected by the founders of contemporary communism have been realized in these really existing societies or not. This is obviously not the case.

See Also

- ▶ Central Planning
- ▶ Collective Agriculture
- ▶ Full Communism
- ▶ Marx, Karl Heinrich (1818–1883)
- ▶ Peasants
- ▶ Planned Economy
- ▶ Socialist Economies

References

Marx, K., and Engels, F. 1845–6. *The German ideology*. As in Karl Marx and Frederick Engels, *Collected works*, vol.5. London: Lawrence & Wishart, 1976.

Community Indifference Curves

Wayne Shafer

Keywords

Bergson–Samuelson social welfare functions; Community indifference curves; Comparative statics; Scitovsky, T.; Utility possibility frontiers

JEL Classifications

D11

The idea of a community indifference curve, as the term is commonly used, is due to Scitovsky (1942). The genesis of the idea is the fact that comparative statics and welfare analysis in economic models is simplified considerably if there is a social preference ordering over aggregate commodity bundles which reflects the collective individual preferences of agents. Scitovksy's notion of a 'community indifference curve' essentially allows the analytical convenience of social indifference curves, in certain circumstances, without having to assume a specific Bergson–Samuelson social welfare function or having to assume the restrictive assumptions on agents' preferences needed to guarantee that agents act collectively as a single individual.

The definition of a community indifference curve is basically simple. Suppose there are m commodities and n agents. Let x denote a commodity vector (as m-vector with non-negative coordinates) and u_i a utility function representing agent i's preferences. We will assume that ui is monotone increasing and quasi-concave. Given a vector $u' = (u_1', \ldots, u_n')$ of utility numbers, the community indifference curve at u', $CIC(u')$, is defined to be the set of all commodity vectors x such that there is a distribution $(x_1, \ldots, x_n)$ of commodity vectors satisfying $\sum_i x_i = x$ and $u_i(x_i) = u_i', i = 1, \ldots, n$, and there is no $x' \leq x$, $x' \neq x$ which also has this property. Thus one can obtain any vector $x \in CIC(u')$ by fixing the quantities of all but one good and minimize the amount of the remaining good subject to achieving u'. As pointed out by Samuelson (1956), the community indifference curve can be interpreted as a 'dual' to the utility possibility frontier. The utility possibility frontier, for a given x, is the set of all vectors u' of utility numbers achievable by a Pareto efficient distribution of x to the agents. Let $U(x)$ denote the utility possibility frontier for the commodity vector x. Then it is easy to see that $CIC(u') = \{x : u' \in U(x)\}$ and that $U(x) = \{u' : x \in CIC(u')\}$.

We will now describe the most important properties of community indifference curves. First, each $CIC(u')$ looks like the indifference curve of a monotone quasi- concave utility function. That is, the set of vectors x such that $x \geq x^1$ for some $x^1 \in CIC(u')$ is a convex set. For example, when $m = 2$, CIC(u') is a curve with a diminishing marginal rate of substitution. Second, unlike the utility possibility frontier, the community indifference curve is essentially an ordinal concept, that is it does not depend on the choice of utility functions representing agents' preferences, in the

following sense. Suppose, for each i, u_i and v_i are two utility functions representing agent i's preferences, and let $(x_1, \ldots, x_n)$ be a Pareto efficient allocation to the agents. Define $u' = [u_1(x_1),\ldots, u_n(x_n)]$ and $v' = [v_1(x_1), \ldots v_n(x_n)]$. Then $CIC(u') \equiv CIV(v')$. Clearly, community indifference curves can be parameterized by a given Pareto efficient allocation of goods rather than a given vector of utilities. Third, assuming smooth utility functions, the marginal rate of substitution for any two commodities on a community indifference curve is equal to the common marginal rate of substitution of each agent. Specifically, pick any $x \in CIC(u')$, and let $(x_1,\ldots,x_n)$ be the Pareto efficient allocation of x such that $u_i(x_i) = u_i^{'}$, $i = 1,\ldots,n$. Then for any two commodities h and h', the marginal rate of substitution of h and h' evaluated at $x \in CIC(u')$ is equal to the marginal rate of substitution of h for h' at x_i on agent i's indifference curve through x_i. Fourth, and very important, community indifference curves are not, in general, 'indifference' curves in the sense of being level curves of some function. Pick any x, and $u^{'}, u^{''} \in U(x)$, such that $u' \neq u^{''}$. Then by definition, $x \in CIC(u') \cap CIC(u^{''})$. Thus $CIC(u')$ must either coincide with $CIC(u^{''})$ or intersect properly. The condition for two community indifference curves never to intersect properly is then that $CIC(u') = CIC(u^{''})$ for all $u', u^{''} \in U(x)$, for all x. It turns out that this is true if and only if the agents have identical homothetic preferences, in which case the family of all community indifference curves will coincide with the family of indifference curves for the common preferences of the agents.

From the above definition and properties, the following observation constitutes the basic use of community indifference curves: if the economy is currently at a vector of utility numbers u', then x' is a commodity vector which lies above $CIC(u')$ if and only if there is some distribution of x' to the agents which will achieve a vector of utilities u'' such that $u^{''} > u'$. In this sense, x' is 'better' than any $x \in CIC(u')$. However, since from above community indifference curves can intersect properly, it may also be that there is a $u^{'''}$ such that $x' \in CIC(u^{'''})$ and an $x' \in CIC(u')$ such that x lies above $CIC(u^{'''})$, in which case x is also 'better' than x'. Thus it is important to realize that community indifference curves cannot be used to define a social ordering of aggregate output vectors. Nevertheless, community indifference curves can still be a useful analytical device. For example, consider a market economy with two produced goods. Consider an equilibrium in which all consumers face the same prices, in terms of the aggregate output vector x' and the vector of utilities u' obtained by the agents. Graphically this equilibrium can be represented by drawing the production possibility frontier and $CIC(u')$, noting they meet at x'. The slope of the production possibility frontier at x' represents the price ratio faced by firms, and the slope of the $CIC(u')$ at x' the common price ratio faced by consumers. If firms and consumers face the same price ratio, then the $CIC(u')$ must be tangent to the production possibility frontier at x. Thus no feasible x can be produced which can make all agents better off, so the situation is Pareto optimal. If, however, firms face different prices than the agents because of, for example, taxes or tariffs, then the slope of the $CIC(u')$ will be different from the slope of the production possibility frontier, and thus the two curves will intersect properly. In this case there must exist an x' on the production possibility frontier which lies above $CIC(u')$, so the original situation is Pareto inefficient.

See Also

- Arrow's Theorem
- Optimality and Efficiency
- Social Welfare Function
- Welfare Economics

Bibliography

Samuelson, P.A. 1956. Social indifference curves. *Quarterly Journal of Economics* 70 (1): 1–22.

Scitovsky, T. 1942. A reconsideration of the theory of tariffs. *Review of Economic Studies* 9 (2): 89–110.

Comparative Advantage

Ronald Findlay

Abstract

This article traces the evolution of the theory of comparative advantage and the gains from trade from the pioneering work of David Ricardo to the factor proportions approach of Eli Heckscher and Bertil Ohlin. Extensions of the basic models to many goods, factors and countries, and to the long run are noted, as well as the attempts at empirical testing of the predictions derived from them.

Keywords

Absolute advantage; Comparative advantage; Corn Laws; Distributive justice; Division of labour; Factor price equalization; Factor proportions; Fixed factors; Free trade; Gains from trade; Haberler, G.; Heckscher, E. F.; Heckscher–Ohlin trade theory; Human capital; International trade; Intra-industry' trade; Labour mobility; Leontief, W.; Lerner, A. P.; Monopolistic competition; Neoclassical general equilibrium theory; Ohlin, B. G.; Product differentiation; Returns to scale; Specialization; Specific factors; Stationary state; Stolper–Samuelson theorem; Terms of trade; Time preference; Wages fund

JEL Classifications

F1

The modern economy, and the very world as we know it today, obviously depends fundamentally on specialization and the division of labour, between individuals, firms and nations. The principle of comparative advantage, first clearly stated and proved by David Ricardo in 1817, is the fundamental analytical explanation of the source of these enormous 'gains from trade'. Though an awareness of the benefits of specialization must go back to the dim mists of antiquity in all civilizations, it was not until Ricardo that this deepest and most beautiful result in all of economics was obtained. Though the logic applies equally to inter*personal,* inter*firm* and inter*regional* trade, it was in the context of inter*national* trade that the principle of comparative advantage was discovered and has been investigated ever since.

The Basic Ricardian Model

What constituted a 'nation' for Ricardo were two things – a 'factor endowment', of a specified number of units of labour in the simplest model, and a 'technology', the productivity of this labour in terms of different goods, such as cloth and wine in his example. Thus labour can move freely between the production of cloth and wine in England and in Portugal, but each labour force is trapped within its own borders. Suppose that a unit of labour in Portugal can produce one unit of cloth or one unit of wine, while in England a unit of labour can produce four units of cloth or two units of wine. Thus the opportunity cost of a unit of wine is one unit of cloth in Portugal while it is two units of cloth in England. On the assumption of competitive markets and free trade, it follows that *both* goods will never be produced in *both* countries since wine in England and cloth in Portugal could always be undermined by a simple arbitrage operation involving export of cloth from England and import of wine from Portugal. Thus wine in England or cloth in Portugal must contract until at least one of these industries produces zero output. If both goods are consumed in positive amounts, the 'terms of trade' in equilibrium must lie in the closed interval between one and two units of cloth per unit of wine. Which of the two countries specializes completely will depend upon the relative size of each country (as measured by the labour force *and* its productivity in each industry) and upon the extent to which each of the two goods is favoured by the pattern of world demand. Thus Portugal is more

likely to specialize the smaller she is compared with England in the sense defined above and the more world demand is skewed towards the consumption of wine relative to the consumption of cloth.

The Gains from Trade

Viewed as a 'positive' theory, the principle of comparative advantage yields *predictions* about (*a*) the *direction* of trade: that each country exports the good in which it has the lower comparative opportunity cost ratio as defined by the technology in that country, and about (*b*) the *terms* of trade: that it is bounded above and below by these comparative cost ratios. From a 'normative' standpoint the principle implies that the citizens of each country become 'better off' as a result of trade, with the extent of the gains from trade depending upon the degree to which the terms of trade exceed the domestic comparative cost ratio. It is the 'normative' part of the doctrine that has always been the more controversial, and it is therefore necessary to evaluate it with the greatest care.

In Ricardo's example the total labour force in each country is presumably supplied by an aggregate of different households, each having the same *relative* productivity in the two sectors. Thus *all* households in *each* country must become better off as a result of trade if the terms of trade lie strictly in between the domestic comparative cost ratios. The import-competing sector in each country simply switches over instantaneously and costlessly to producing the export good (moving to the opposite corner of its linear production-possibilities frontier, in terms of the familiar geometry), obtaining the desired level of the other good by imports, raising utility in the process. When one country is incompletely specialized, then all households in that country remain at unchanged utility levels, all of the gain from trade going to the individuals in the 'small' country. Thus we have a situation in which *everybody gains,* in at least one country, while *nobody loses* in either country, as a result of trade.

This very strong result depends upon Ricardo's assumption of perfect occupational mobility in each country. Suppose we take the opposite extreme of completely *specific* labour in each sector, so that each country produces a fixed combination of cloth and wine, with no possibility of transformation. In this case, labour in the import-competing sector in each country must necessarily *lose,* as a result of trade, while labour in each country's export sector must gain. It can be shown, however, that trade will improve *potential* welfare in each country in the Samuelson (1950) sense that the utility-possibility frontier with trade will dominate the corresponding frontier without trade, so that no one need be worse off, and at least some one better off, if lump-sum taxes and transfers are possible (Samuelson 1962).

International Factor Mobility and World Welfare

Another very important normative issue is the question of the relationship between the free-trade equilibrium and *world* efficiency and welfare. In the Ricardian model world welfare in general will *not* be maximized by free trade alone. In the numerical example considered here Ricardo stresses the fact that England can still gain from trade even though she has an *absolute* advantage in the production of *both* goods, her productivity being greater in both cloth and wine, though comparatively greater in cloth. Suppose that labour in Portugal could produce at English levels, *if it moved to England*; that is, the English superiority is based on climatic or other 'environmental' factors and not on differences in aptitude or skill. Then, if labour was free to move, and in the absence of 'national' sentiment, all production would be located in England, and Portugal would cease to exist. The former Portuguese labour would be better off than under free trade, since their real wage in terms of wine will now be two units instead of one. The English labour would be worse off, if the terms of trade were originally better than 0.5 wine per unit of cloth, but it is easy to show that they could be sufficiently compensated since the utility-possibility frontier for the world economy as a whole is moved out by the integration of the labour forces.

The case when each country has an absolute advantage in one good is more interesting. As is easy to see, from Findlay (1982), this case will involve a movement of labour to the country with the higher real wage under free trade, increasing the production of this country's exportable and reducing that of the lower-wage country under free trade. The terms of trade turn against the higher-wage country until eventually the real wage is equalized. The terms of trade that achieve this equality of real wages will be equal to the ratio of labour productivities in each country's export sector; that is, the 'double factoral' terms of trade will be unity. This solution of free trade *combined* with perfect labour mobility will achieve not only efficiency for the world economy as a whole but equity as well. 'Unequal exchange' in the sense of Emmanuel (1972) would not exist, while liberal, utilitarian and Rawlsian criteria of distributive justice would be satisfied as well, as pointed out in Findlay (1982). Despite all this, it still seems utopian to expect a policy of 'open borders', in *either* direction, for the contemporary world of nation-states.

Extensions of the Basic Ricardian Model

The two-country, two-good Ricardian model was extended to many goods and countries by a number of subsequent writers, whose efforts are described in detail by Haberler (1933) and Viner (1937). In the case of two countries and n goods the concept of a 'chain of comparative advantage' has been put forward, with the goods listed in descending order in terms of the *relative* efficiency of the two countries in producing them. It is readily shown that with a uniform wage in each country all goods from 1 to some number j must be exported, while all goods from $(j + 1)$ to n must be imported. The number j itself will depend upon the relative sizes of the two countries and the composition of world demand. Dornbusch et al. (1977) generalize this result to a continuum of goods in an extremely elegant and powerful model that has been widely used in subsequent literature. An analogous chain concept applies to the case of two goods and n countries, this time ranking the countries in terms of the ratio of their productivities in the two goods, with country 1 having the greatest *relative* efficiency in cloth and country n in wine. World demand and the sizes of the labour forces will determine the 'marginal' country j, with countries 1 to j exporting cloth and $(j + 1)$ to n exporting wine.

The simultaneous consideration of comparative advantage with many goods and many countries presents severe analytical difficulties. Graham (1948) considered several elaborate numerical examples, his work inspiring the Rochester theorists McKenzie (1954) and Jones (1961) to apply the powerful tools of activity analysis to this particular case of a linear general equilibrium model. It is interesting to note in connection with mathematical programming and activity analysis that Kantorovich (1965) in his celebrated book on planning for the Soviet economy worked out an example of optimal specialization patterns for factories that corresponds *exactly* to the Ricardian model of trade between countries.

The Three-Factor Ricardian Model

While most of the literature on the Ricardian trade model has concentrated on the model of Chapter 7 of the *Principles* in which it appears that labour is the sole scarce factor, his more extended model in the *Essay on Profits* has been curiously neglected, though the connections between trade, income distribution and growth which that analysis explores are quite fascinating. The formal structure of the model was laid out very thoroughly in Pasinetti (1960). The economy produces two goods, corn and manufactures, each of which has a one-period lag between the input of labour and the emergence of output. Labour thus has to be supported by a 'wage fund', an initially given stock that is accumulated over time by saving out of profits. Corn also requires land as an input, which is in fixed supply and yields diminishing returns to successive increments of labour. The wage-rate is given exogenously in terms of corn, and manufactures are a luxury good consumed only by the land-owning class, who obtain rents determined by the marginal product of land.

Profits are the difference between the marginal product of labour and the given real wage, which is equal to the marginal product 'discounted' by the rate of interest, in this model equal to the rate of profit, defined as the ratio of profits to the real wage that has to be advanced a period before. Momentary equilibrium determines the relative price of corn and manufactures, the rent per acre and the rate of profit, as well as the output levels and allocation of the labour force between sectors. The growth of the system is at a rate equal to the product of the rate of profit and the propensity to save of the capitalist class. It is shown that the system approaches a stationary state, with a monotonically falling rate of profit and rising rents per acre.

The opportunity to import corn more cheaply from abroad will have significant distributional and growth consequences. Just as Ricardo argued in his case for the repeal of the Corn Laws, cheaper foreign corn will reduce domestic rents and raise the domestic rate of profit, and thus the rate of growth. The approach to the stationary state is postponed, though of course it cannot be ultimately averted, while the growth consequences for the corn exporter are definitely adverse. The main doctrinal significance of this wider Ricardian model, however, is to reveal the extent to which the subsequent 'general equilibrium' or 'neoclassical' approach to international trade is already present within the Ricardian framework. For one thing, the pattern of comparative advantage itself depends upon the complex interaction of technology, factor proportions and tastes. In his Chapter 7 case the pattern of comparative advantage is *exogenous,* simply given by the four fixed technical coefficients indicating the productivity of labour in cloth and wine in England and Portugal. The production-possibility frontiers for each country are linear, and comparative advantage is simply determined by the relative magnitudes of the slopes. As demonstrated in Findlay (1974), however, the *Essay on Profits* model implies a concave production-possibilities frontier at any moment, since there are diminishing returns to labour in corn even though the marginal productivity of labour in manufactures is constant. With two countries the pattern of comparative advantage will depend upon the slopes of these curves at their autarky equilibria, which are *endogenous* variables depending upon the sizes of the 'wage fund' in relation to the supply of land and the consumption pattern of landowners, as well as the technology for the two goods.

As Burgstaller (1986) points out, however, the steady-state solution of the model restores the linear structure of the pattern of comparative advantage. The zero profit rate in the steady state requires the marginal product of labour to be equal to the given real wage, and this implies a fixed land–labour ratio and hence output per unit of labour in corn. We thus once again have two fixed technical coefficients, so that the slope of the linear production-possibilities frontier is once again an exogenous indicator of comparative advantage.

The 'neo-Ricardian' approach of Steedman (1979a, b) considers more general time-phased structures of production. Technology alone determines negatively sloped wage–profit or factor-price frontiers, any point on which generates a set of relative product prices and hence a pattern of comparative advantage relative to another such economy.

Factor Proportions and the Heckscher–Ohlin Model

While J.S. Mill, Marshall and Edgeworth all made major contributions to trade theory, the concept of comparative advantage did not undergo any evolution in their work beyond the stage at which Ricardo had left it. They essentially concentrated on the determination of the terms of trade and on various comparative static exercises. The interwar years, however, brought fundamental advances, stemming in particular from the work of the Swedes Heckscher (1919) and Ohlin (1933). The development of a diagrammatic apparatus to handle general equilibrium interactions of tastes, technology and factor endowments by Haberler (1933), Leontief (1933), Lerner (1932) and others culminated in the rigorous establishment of trade theory and comparative advantage as a branch of neoclassical general equilibrium theory.

The essentials of this approach can be expounded in terms of the familiar two-country, two-good and two-factor model, on which see Jones (1965) for a detailed and lucid algebraic exposition. The given factor supplies and constant returns to scale technology define concave production-possibility frontiers, on the assumption that the goods differ in factor intensity. This determines the 'supply side' of the model, which is closed by the specification of consumer preferences. Economies that have identical technology, factor endowments and tastes will have the same autarky equilibrium price-ratio and so will have no incentive to engage in trade. Countries must therefore differ with respect to at least one of these characteristics for differences in comparative advantage to emerge. With identical technology and factor endowments, a country will have a comparative advantage in the good its citizens prefer *less* in comparison to the foreign country, since then this good will be cheaper at home. Similarly, if factor endowments and tastes are identical, differences in comparative advantage will be governed by relative technological efficiency; that is, a country will have a comparative advantage in the good in which its relative technological efficiency is greater, just as in the Ricardian model. These differences in technological efficiency could be represented, for example, by the magnitude of multiplicative constants in the production functions; that is, 'Hicks-neutral' differences.

In keeping with the ideas of Heckscher and Ohlin, however, it is differences in factor proportions that have dominated the explanation of comparative advantage in the neoclassical literature. The Heckscher–Ohlin theorem, that each country will export the commodity that uses its relatively abundant factor most intensively, has been rigorously established and the necessary qualifications carefully specified, as in Jones (1956). Among the more important of these is the requirement that factor-intensity 'reversals' do not take place; that is, that one good is always more capital-intensive than the other at all wage-rental ratios or at least within the relevant range defined by the factor proportions of the trading countries.

The Stolper–Samuelson Theorem

Associated with the Heckscher–Ohlin theorem is the Stolper–Samuelson theorem (1941), that trade benefits the abundant and harms the scarce factor while protection does the opposite, and the celebrated factor price equalization theorem of Lerner (1952, though written in 1932) and Samuelson (1948, 1949, 1953), which states that under certain conditions free trade will lead to complete equalization of factor rewards even though factors are not mobile internationally. The normative significance of this theorem is that free trade alone can achieve world efficiency in production and resource allocation, unlike the case of the Ricardian model as pointed out earlier. The requirements for the theorem to hold, however, are very stringent, such as that the number of tradable goods produced be equal to the number of factors. It also requires factor proportions to be sufficiently close to each other in the trading partners so that the production patterns are fairly similar. Thus it would be far-fetched to expect the price of unskilled labour to be equalized between Bangladesh and the United States, for example.

The Specific-Factors Model

An important and popular variant of the factor proportions approach is what Jones (1971) calls the 'specific factors' and Samuelson (1971) the Viner–Ricardo model. In this model each production sector has its own unique fixed factor, while labour is used in all sectors and is perfectly mobile internally between them. Trade patterns still reflect factor endowments but factor price equalization does not hold in this model since the number of factors is always one greater than the number of goods. Gruen and Corden (1970) present an ingenious three-by-three extension of this approach, in which one sector uses land and labour, while the two others use capital and labour, thus neatly integrating the 'specific factors' model with the regular two-by-two Heckscher–Ohlin model. Findlay (1995, chs. 4 and 6) uses adaptations and extensions of the Gruen–Corden specification to introduce human

capital formation and the concept of a natural resource 'frontier' into the Heckscher–Ohlin framework.

Long-Run Extensions of the Factor Proportions Model

One limitation of the Heckscher–Ohlin model was that the stock of 'capital', however conceived, should be an endogenous variable determined by the propensity to save or time preference of each trading community, rather than being taken as exogenously fixed. Oniki and Uzawa (1965) extended the model to a situation where the labour force is growing in each country at an exogenous rate and capital is accumulated in response to given propensities to save in each country. One of the goods is taken to be the 'capital' good, conceived of as a malleable 'putty–putty' instrument. They demonstrated that the system converges in the long run to a particular capital–labour ratio for each country, which will be higher for the country with the larger saving propensity. In Findlay (1970, 1984), it is shown that as the capital–labour ratio evolves the pattern of comparative advantage for a given 'small' country in an open trading world will also shift over time towards more capital-intensive goods, thus formalizing the concept of a 'ladder of comparative advantage' that countries ascend in the process of economic development. Thus comparative advantage should not be conceived as given and immutable, but evolving with capital accumulation and technological change. Much of the loose talk about 'dynamic' comparative advantage in the development literature, however, is misconceived since it attempts to change the pattern of production by protection *before* the necessary changes in the capacity to produce efficiently have taken place. Other models which endogenize the capital stocks of the trading countries are Stiglitz (1970) and Findlay (1978) which utilizes a variable rate of time preference and an 'Austrian' point-input/point-output technology, which implies a continuum of capital goods as represented by the 'trees' of different ages, and Findlay (1995, ch. 2), which addresses the question posed by Samuelson (1965) of whether trade equalizes not only the marginal product or rental of capital but the rate of interest itself.

Empirical Testing

Empirical testing of the positive side of the theory of comparative advantage begins in a systematic way only with the work of MacDougall (1951) on the Ricardian theory and the celebrated article of Leontief (1954) that uncovered the apparent paradox that US exports were more labour-intensive than her imports. Leontief's dramatic finding spurred considerable further empirical research motivated by the desire to find a satisfactory explanation. The increasing scarcity of natural resources in the USA, by causing capital to be substituted for it in import-competing production, was stressed as an explanation for the paradox by Vanek (1963). The role of 'human capital' as an explanation was pointed to by Kenen (1965) and a number of empirical investigators, who found that US exports were considerably more skill-intensive than her imports, even though physical capital-intensity was only weakly correlated with exports and imports. This pointed to the need to reinterpret the simple Heckscher–Ohlin model in terms of skilled and unskilled labour as the two factors, rather than labour of uniform quality and physical capital. Since the formation of skill through education is an endogenous variable, a function of a wage differential that is itself a function of trade, we need a general equilibrium model that can simultaneously handle both these aspects, as in Findlay and Kierzkowski (1983) and Findlay (1995, ch. 4).

Many other extensions of the Heckscher–Ohlin theory are surveyed in Jones and Neary (1984) and Ethier (1984), while Deardorf (1984) and Feenstra (2004) give very incisive accounts of the attempts at empirical testing of the theory of comparative advantage in its different manifestations. Further important progress in empirical testing of the Heckscher–Ohlin model has been achieved by the work of Leamer (1984), Trefler (1995), Harrigan (1997) and Davis and Weinstein (2001).

Increasing Returns

Finally, the crucial role of increasing returns to scale in specialization and international trade has only recently been rigorously investigated, since it implies departures from perfect competition. Krugman (1979) and Lancaster (1980) introduced international trade into models of monopolistic competition with differentiated products, showing the possibility of gains from trade due to the provision of greater variety of similar goods rather than differences in comparative advantage, what is referred to as 'intra-industry' trade. Helpman and Krugman (1985) thoroughly examine and extend our knowledge in this area, while Grossman and Helpman (1991) expertly extend the monopolistic competition approach to deal with a number of issues involving endogenous technological progress and growth in the world economy.

See Also

- ▶ Heckscher, Eli Filip (1879–1952)
- ▶ Heckscher–Ohlin Trade Theory
- ▶ International Trade Theory
- ▶ Leontief, Wassily (1906–1999)
- ▶ Ohlin, Bertil Gotthard (1899–1979)
- ▶ Ricardo, David (1772–1823)
- ▶ Terms of Trade

Bibliography

Burgstaller, A. 1986. Unifying Ricardo's theories of growth and comparative advantage. *Economica* 53: 467–481.

Davis, D.R., and D.E. Weinstein. 2001. An account of global factor trade. *American Economic Review* 91: 1423–1453.

Deardorf, A. 1984. Testing trade theories. In *Handbook of international economics*, ed. R.W. Jones and P.B. Kenen, vol. 1. Amsterdam: North-Holland.

Dombusch, R., S. Fischer, and P.A. Samuelson. 1977. Comparative advantage, trade and payments in a Ricardian model with a continuum of goods. *American Economic Review* 67: 823–839.

Emmanuel, A. 1972. *Unequal exchange*. New York: Monthly Review Press.

Ethier, W. 1984. Higher dimensional issues in trade theory. In *Handbook of international economics*, ed. R.W. Jones and P.B. Kenen, vol. 1. Amsterdam: North-Holland.

Feenstra, R.C. 2004. *Advanced international trade*. Princeton: Princeton University Press.

Findlay, R. 1970. Factor proportions and comparative advantage in the long run. *Journal of Political Economy* 78: 27–34.

Findlay, R. 1974. Relative prices, growth and trade in a simple Ricardian system. *Economica* 41: 1–13.

Findlay, R. 1978. An 'Austrian' model of international trade and interest equalization. *Journal of Political Economy* 86: 989–1007.

Findlay, R. 1982. International distributive justice. *Journal of International Economics* 13: 1–14.

Findlay, R. 1984. Growth and development in trade models. In *Handbook of international economics*, ed. R.W. Jones and P.B. Kenen, vol. 1. Amsterdam: North-Holland.

Findlay, R. 1995. *Factor proportions, trade, and growth*. Cambridge, MA: MIT Press.

Findlay, R., and H. Kierzkowski. 1983. International trade and human capital: A simple general equilibrium model. *Journal of Political Economy* 91: 957–978.

Graham, F. 1948. *The theory of international values*. Princeton: Princeton University Press.

Grossman, G.M., and E. Helpman. 1991. *Innovation and growth in the global economy*. Cambridge, MA: MIT Press.

Gruen, F., and W.M. Corden. 1970. A tariff that worsens the terms of trade. In *Studies in international economics*, ed. I.A. MacDougall and R. Snape. Amsterdam: North-Holland.

Haberler, G. 1933. *The theory of international trade*. Trans. A. Stonier and F. Benham. London: W. Hodge, 1936; revised edn, 1937.

Harrigan, J. 1997. Technology, factor supplies and international specialization: Estimating the neoclassical model. *American Economic Review* 87: 475–494.

Heckscher, E. 1919. The effects of foreign trade on the distribution of income. In *Ekonomisk Tidskrift*. English translation in *Readings in the Theory of International Trade,* ed. H.S. Ellis and L.A. Metzler. Philadelphia: Blakiston, 1949.

Helpman, E., and P. Krugman. 1985. *Market structure and foreign trade*. Cambridge, MA: MIT Press.

Jones, R.W. 1956. Factor proportions and the Heckscher–Ohlin theorem. *Review of Economic Studies* 24: 1–10.

Jones, R.W. 1961. Comparative advantage and the theory of tariffs. *Review of Economic Studies* 28: 161–175.

Jones, R.W. 1965. The structure of simple general equilibrium models. *Journal of Political Economy* 73: 557–572.

Jones, R.W. 1971. A three-factor model in theory, trade and history. In *Trade, balance of payments, and growth*, ed. J. Bhagwati et al. Amsterdam: North-Holland.

Jones, R.W., and P. Neary. 1984. Positive trade theory. In *Handbook of international economics*, ed. R.W. Jones and P.B. Kenen, vol. 1. Amsterdam: North-Holland.
Kantorovich, L. 1965. *The best use of economic resources*. Cambridge, MA: Harvard University Press.
Kenen, P.B. 1965. Nature, capital and trade. *Journal of Political Economy* 73: 437–460; Erratum, December, 658.
Krugman, P.R. 1979. Increasing returns, monopolistic competition and international trade. *Journal of International Economics* 9: 469–479.
Lancaster, K.J. 1980. Intra-industry trade under perfect monopolistic competition. *Journal of International Economics* 10: 151–175.
Leamer, E.P. 1984. *Sources of international comparative advantage: Theory and evidence*. Cambridge, MA: MIT Press.
Lemer, A.P. 1952. Factor prices and international trade. *Economica* 19: 1–15.
Leontief, W.W. 1933. The use of indifference curves in the analysis of foreign trade. *Quarterly Journal of Economics* 47: 493–503.
Leontief, W.W. 1954. Domestic production and foreign trade: The American capital position re-examined. *Economia Internazionale* 7: 9–38.
Lerner, A.P. 1932. The diagrammatic representation of cost conditions in international trade. *Economica* 12: 345–356.
Lerner, A.P. 1952. Factor prices and international trade. *Economica* 19: 1–15.
MacDougall, G.D.A. 1951. British and American exports. *Economic Journal* 61: 697–724.
McKenzie, L.W. 1954. Specialization and efficiency in world production. *Review of Economic Studies* 21 (3): 165–180.
Ohlin, B. 1933. *Inter-regional and international trade*. Cambridge, MA: Harvard University Press.
Oniki, H., and H. Uzawa. 1965. Patterns of trade and investment in a dynamic model of international trade. *Review of Economic Studies* 32: 15–38.
Pasinetti, L. 1960. A mathematical formulation of the Ricardian system. *Review of Economic Studies* 27: 78–98.
Ricardo, D. 1951. *The works and correspondence of David Ricardo*, ed. P. Sraffa. vols. 1 and 4. Cambridge: Cambridge University Press.
Samuelson, P.A. 1948. International trade and the equalization of factor prices. *Economic Journal* 58: 163–184.
Samuelson, P.A. 1949. International factor price equalization once again. *Economic Journal* 59: 181–197.
Samuelson, P.A. 1950. Evaluation of real national income. *Oxford Economic Papers* 2: 1–29.
Samuelson, P.A. 1953. Prices of factors and goods in general equilibrium. *Review of Economic Studies* 21: 1–20.
Samuelson, P.A. 1962. The gains from international trade once again. *Economic Journal* 72: 820–829.
Samuelson, P.A. 1965. Equalization by trade of the interest rate along with the real wage. In *Trade, growth and the balance of payments*, ed. R. Baldwin et al. Chicago: Rand McNally.
Samuelson, P.A. 1971. Ohlin was right. *Swedish Journal of Economics* 73: 365–384.
Steedman, I. 1979a. *Trade amongst growing economies*. Cambridge: Cambridge University Press.
Steedman, I., ed. 1979b. *Fundamental issues in trade theory*. London: Macmillan.
Stiglitz, J. 1970. Factor–price equalization in a dynamic economy. *Journal of Political Economy* 78: 456–488.
Stolper, W., and P.A. Samuelson. 1941. Protection and real wages. *Review of Economic Studies* 9: 58–73.
Trefler, D. 1995. The case of missing trade and other mysteries. *American Economic Review* 85: 1029–1046.
Vanek, J. 1963. *The natural resource content of U.-S. Foreign trade 1870–1955*. Cambridge, MA: MIT Press.
Viner, J. 1937. *Studies in the theory of international trade*. New York: Harper.

Comparative Statics

John Nachbar

Abstract

Comparative statics in competitive general equilibrium (GE) environments provide insight into the operation of GE models and a means to confront GE models with data. This article focuses on a canonical comparative statics prediction: in an exchange economy, aggregate endowment changes are negatively related to equilibrium price changes. In particular, an increase in the aggregate endowment of a commodity lowers its equilibrium price.

Keywords

Aggregate excess demand; Asset pricing; Comparative statics; Competitive general equilibrium; Debreu–Mantel–Sonnenschein theorem; Demand shocks; Endowment changes; General equilibrium; Gross substitutes; Hicks, J.; Partial equilibrium; Production economies; Rybczynski theorem; Stolper–Samuelson theorem; Supply shocks; Weak axiom of revealed preference

JEL Classifications
C6

Comparative statics in competitive general equilibrium (GE) environments provide insight into the operation of GE models and a means, at least in principle, to confront GE models with data.

For concreteness, I focus most of this article on what is arguably the canonical GE comparative statics conjecture: in finite exchange economies (that is, no production), equilibrium price changes are negatively related to endowment changes. In particular, if the endowment of good 1 increases and the endowments of other goods remain the same, then the price of good 1 falls. At the end of this article, I briefly survey other GE comparative statics results.

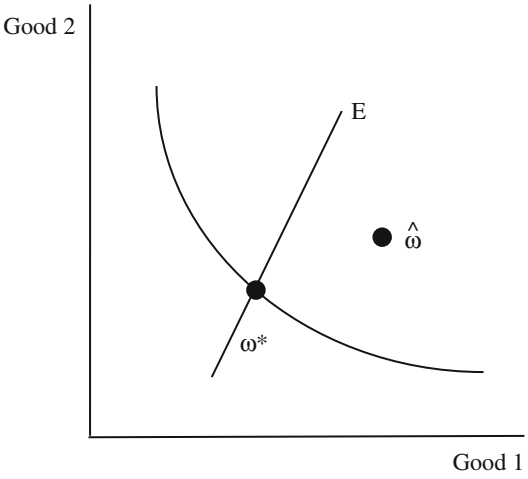
I break the analysis into three cases of increasing complexity.

Case I

There is a single consumer and two commodities. In an equilibrium of this trivial economy, the consumer eats her own endowment. Equilibrium relative prices (which are well defined even though there is no trade) are given by the slope of the consumer's indifference curve through her consumption/endowment point, ω^*. Let E denote her wealth expansion path through her initial endowment; E is the set of points where the slope of her indifference curve is the same as at ω^*.

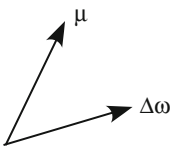
If the new endowment, $\hat{\omega}$ lies below E, as in Fig. 1, then the equilibrium price ratio p_1/p_2 falls. If $\hat{\omega}$ lies above E, then p_1/p_2 rises. If $\hat{\omega}$ lies along E, then p_1/p_2 remains unchanged.

The differential version of Fig. 1 is given by Fig. 2. The vector μ, the tangent to E, is the derivative with respect to nominal wealth of each good's demand (μ is mnemonic for marginal propensity to consume vector). To first order, the rule is that p_1/p_2 falls if and only if $\Delta\omega$, the vector of endowment changes, lies within 180° clockwise from μ. The vector μ incorporates second order information from the utility function and is, in particular, typically not collinear with the utility gradient.



Comparative Statics, Fig. 1 Comparative statics with one consumer and two commodities

Comparative Statics, Fig. 2 First order comparative statics with one consumer and two commodities



If the endowment of good 1 increases while the endowment of good 2 remains unchanged, then $\Delta\omega$ lies along the positive good 1 axis. Figure 2 implies that, in this case, p_1/p_2 falls if and only if good 2 is normal ($\mu_2 > 0$); whether good 1 is normal or inferior (or even a Giffen good) is irrelevant.

Case II

There is again one consumer but L commodities. If $\Delta\omega$ lies along the positive good 1 axis, then a natural conjecture, to generalize the above observation for $L = 2$, is that p_1/p_ℓ falls for each good $\ell > 1$, provided each of these goods is normal. Hicks (1939) showed that this conjecture is false in general but that it is true if the gross substitute property holds (GS; the matrix of partial derivatives of excess demand with respect to price has positive off-diagonal entries). GS holds automatically in the one-consumer, $L = 2$, case because, at equilibrium, when $L = 2$, GS is equivalent to the weak axiom of revealed preference (WA).

Matters are more complicated if two or more endowments are shifting at the same time. For multivariate endowment shocks, there appears to be little one can say in general about changes in the price ratio of any specific pair of commodities. Instead, the conjecture is that $\Delta\omega$ is negatively related to Δp, the vector of equilibrium price changes. Formally,

$$\Delta p \cdot \Delta\omega \leq 0.$$

Call this the comparative statics inequality, CS for short. Geometrically, CS says that the vectors Δp and $\Delta\omega$ lie at least 90° apart.

To interpret Δp as a change in relative prices, prices must be normalized. Consider linear price normalizations, in which all prices, in both the original economy and the perturbed economy, satisfy $p \cdot \lambda = 1$, where λ is an L vector. If all the coordinates of λ are positive, then a fall in the normalized price of good 1 means that the ratio

$$\frac{p_1}{p_{-1} \cdot \lambda_{-1}}$$

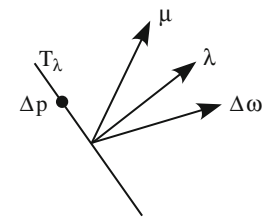
falls, where p_{-1} and λ_{-1} are subvectors corresponding to all goods other than the first: the price of good 1 falls relative to the value of a commodity bundle consisting of λ_ℓ units of each good $\ell > 1$. Standard choices of λ include $\lambda = (0, \ldots, 0, 1)$ (use the last commodity as numeraire) and $\lambda = \omega^*$ (normalize prices so that GNP remains constant; this is the Laspeyres normalization). Regardless of how, or whether, actual prices are normalized, one can re-normalize prices using whatever λ one chooses.

I can provide intuition for CS most easily by continuing to use figures for a two-good economy. Fix a normalizing vector λ Since $p \cdot \lambda = 1$ for all p, $\Delta p \cdot \lambda = 0$. Therefore, Δp lies along the line that is at right angles to λ, labelled T_λ in Fig. 3.

As drawn, $\Delta\omega$ lies within 180° clockwise from μ and hence p_1/p_2 falls. Therefore, Δp, normalized by λ, lies on the upper left-hand branch of T_λ. As illustrated, $\Delta\omega$ and Δp are more than 90° apart; hence CS holds.

On the other hand, suppose that $\Delta\omega$ lies in the cone spanned by λ and μ. Since $\Delta\omega$ again lies within 180° clockwise from μ, p_1/p_2 again falls and Δp again lies on the upper left-hand branch of T_λ. Now, however, $\Delta\omega$ and Δp are less than 90° apart. CS fails.

Comparative Statics, Fig. 3 Condition CS with two goods



In general, in a one-consumer economy, for any number of commodities and for any preferences, CS fails whenever there is a gap between λ and μ and $\Delta\omega$ falls into this gap. Conversely, if $\lambda = \mu$ (or, more generally, if λ is a scalar multiple of μ) then CS holds for any endowment change: $\Delta p \cdot \Delta\omega \leq 0$ with $\Delta p \cdot \Delta\omega = 0$ if and only if $\Delta\omega$ is collinear with μ (which is the differential analog of $\Delta\omega$ landing on the wealth expansion path E in Fig. 1). In one consumer economies, $\lambda = \mu$ is thus the unique (up to scalar multiplication) linear price normalization for which CS holds for all endowment changes.

If preferences are quasi-linear in good L, and consumption is interior, then $\lambda = \mu$ implies $\lambda = (0, \ldots, 0, 1)$; the last good is used as numeraire. If the preferences are homothetic then μ is a scalar multiple of the reference endowment, ω^*, and so one can set $\lambda = \omega^*$. Typically, however, $\lambda = \mu$ is different from price normalizations commonly used in economics.

The $\lambda = \mu$ normalization, although nonstandard, does have a sensible interpretation, provided μ is positive (all goods are weakly normal). If μ is positive, then a decrease in p_1 means that $p_1/(p_{-1} \cdot \mu_{-1})$ falls: the price of good 1 falls relative to the value of the consumer's *marginal* consumption of all other goods.

If $\Delta\omega$ lies along the positive good 1 axis and goods 2, . . ., L are normal, then a minor variation on CS implies that $p_1/(p_{-1} \cdot \mu_{-1})$ falls, even if good 1 is inferior. This is a Weaker conclusion then that of Hicks (1939) but it has a weaker hypothesis, since it does not assume the gross substitute property.

Case III

There are I consumers and L commodities. The generalization of CS is

$$\Delta p \cdot \Delta \overline{\omega} \leq 0,$$

where $\Delta \overline{\omega}$ denotes the change in the aggregate endowment. CS holds provided one uses an appropriate aggregate version of μ. Consider two alternatives. Each is a weighted sum of the individual marginal propensity to consume vectors, μ^i. In the first, $\overline{\mu}_{\Delta x}$, the weight on μ^i is consumer i's share in the change in the value of consumption, evaluated at the prices of the reference equilibrium. In the second, $\overline{\mu}_{\Delta \omega}$, the weight on μ^i is i's share in the change in the value of the endowment, again evaluated at reference equilibrium prices.

If one normalizes prices using $\lambda = \overline{\mu}_{\Delta x}$, then inequality CS holds provided *individual* excess demand satisfies the weak axiom (WA) at equilibrium. If $\lambda = \overline{\mu}_{\Delta \omega}$, then CS holds provided *aggregate* excess demand satisfies WA at equilibrium. See Nachbar (2002).

The hypothesis that aggregate excess demand satisfies WA is not implied by standard GE assumptions. One justification for nevertheless assuming WA is that it seems to be connected to the dynamic stability of the price adjustment process. WA holding at equilibrium is sufficient and almost necessary for local asymptotic stability under the Walrasian tâtonnement, for example. Comparative statics, by assuming that economies are at equilibrium, may therefore implicitly assume that aggregate excess demand satisfies WA. A second justification for assuming that aggregate excess demand satisfies WA is that this assumption, while strong, is not implausible in exchange economies. For some sufficient conditions for WA, see Becker (1962), Hildenbrand (1983), Grandmont (1992) and Quah (1997).

In the one-consumer case, the $\lambda = \mu$ price normalization was necessary as well as sufficient. There are analogous, but weaker, necessity results for $\overline{\mu}_{\Delta x}$ and $\overline{\mu}_{\Delta \omega}$. The important implication is that, because both $\overline{\mu}_{\Delta x}$ and $\overline{\mu}_{\Delta \omega}$ can vary with how the endowment changes are distributed across consumers, there may be *no* price normalization for which CS holds for all endowment changes. As the endowment distribution changes, the price normalization may have to change.

This illustrates an issue that has become a central theme in the recent literature on GE comparative statics. Given an arbitrary price normalization, standard GE assumptions impose no restrictions on the relationship between changes in equilibrium prices and changes in the aggregate endowment (see Chiappori et al. 2004). This negative result, a cousin of the Debreu–Mantel–Sonnenschein theorem (DMS), has a loophole: standard GE assumptions do provide comparative statics restrictions if one works with micro-level information (for example, on the endowment distribution) rather than exclusively with aggregates. In the CS results, micro-level data is used in the price normalization. Note that CS requires micro data even if one assumes that aggregate excess demand satisfies WA.

Relative to the objectives laid out in the first paragraph of this article, the results on CS comparative statics fare reasonably well in providing insight into the operation of GE models. The $\overline{\mu}_{\Delta \omega}$ result is much the easier to interpret, since it is computed with the use of endowment changes, which are exogenous. The $\overline{\mu}_{\Delta x}$ result, on the other hand, extends easily to production economies. In contrast, I do not know whether the $\overline{\mu}_{\Delta \omega}$ result has a useful analog for production economies.

The CS inequality fares less well as a tool for empirical work, because it requires a large amount of data just to estimate the normalization vector. The necessity results imply that this difficulty is intrinsic to CS.

Other Comparative Statics Results

Brown and Matzkin (1996), a path-breaking paper that has heavily influenced subsequent work in this area, exploits the DMS loophole noted above to give testable restrictions linking equilibrium prices with individual endowments. For related work, see Snyder (1999), Williams (2002), Kübler (2003) and Chiappori et al. (2004). Relative to CS, the Brown–Matzkin restrictions are easier to implement empirically

because they do not require estimating normalization vectors, but they are harder to interpret.

As already noted, CS-type reasoning can be extended to production economies (see Quah 2003; Nachbar 2004). CS-type reasoning can also be extended to asset pricing environments (see Quah 2003).

For shocks to preferences rather than endowments or technologies, the analog of CS is

$$\Delta p \cdot \Delta \overline{x} \leq 0,$$

where $\Delta \overline{x}$ is the change in equilibrium consumption. Profit maximization implies that this inequality holds for any price normalization. In this respect, the analysis of demand shocks is trivial compared with the analysis of supply shocks.

Interest in comparative statics has helped motivate research on the uniqueness, regularity, and stability of equilibria (see Kehoe 1987). Note that some of the comparative statics results cited above (for example, the Brown–Matzkin results and the $\overline{\mu}_{\Delta x}$ CS result) do not assume uniqueness or stability.

Finally, perhaps the most famous comparative statics results are the Stolper-Samuelson theorem and its dual, the Rybcyznski theorem (for a recent treatment, see Echenique and Manelli 2005). Stolper–Samuelson links changes in factor prices with factor intensities and changes in output prices. Rybcyznski links changes in final goods production with factor intensities and changes in factor supplies. Although it is possible to embed these results within a highly restricted GE model, they are partial equilibrium in spirit; wealth effects play no role.

See Also

- ▶ General Equilibrium
- ▶ General Equilibrium (New Developments)
- ▶ International Trade Theory

Bibliography

Becker, G. 1962. Irrational behavior and economic theory. *Journal of Political Economy* 70: 1–13.

Brown, D., and R. Matzkin. 1996. Testable restrictions on the equilibrium manifold. *Econometrica* 64: 1249–1262.

Chiappori, P., I. Ekelund, F. Kübler, and H. Polemarchakis. 2004. Testable implications of general equilibrium theory: A differentiable approach. *Journal of Mathematical Economics* 40: 105–119.

Echenique, F., and A. Manelli. 2005. *Comparative statics, English auctions, and the Stolper–Samuelson theorem*. Mimeo. Tempe: Arizona State University.

Grandmont, J. 1992. Transformations of the commodity space, behavioral heterogeneity and the aggregation problem. *Journal of Economic Theory* 57: 1–35.

Hicks, J. 1939. *Value and capital*. Oxford: Clarendon Press.

Hildenbrand, W. 1983. On the law of demand. *Econometrica* 51: 997–1018.

Kehoe, T. 1987. Comparative statics. In *The new Palgrave: A dictionary of economics*, ed. J. Eatwell, M. Milgate, and P. Newman, vol. 1. Basingstoke: Palgrave.

Kübler, F. 2003. Observable restrictions of general equilibrium with financial markets. *Journal of Economic Theory* 110: 137–153.

Nachbar, J. 2002. General equilibrium comparative statics. *Econometrica* 79: 2065–2074.

Nachbar, J. 2004. General equilibrium comparative statics: The discrete case with production. *Journal of Mathematical Economics* 40: 153–163.

Quah, J. 1997. The law of demand when income is price dependent. *Econometrica* 65: 1421–1442.

Quah, J. 2003. Market demand and comparative statics when goods are normal. *Journal of Mathematical Economics* 39: 317–333.

Snyder, S. 1999. Testable restrictions of Pareto optimal public good provision. *Journal of Public Economics* 71: 97–119.

Williams, S. 2002. *Equations on the derivatives of an initial endowment-competitive equilibrium mapping for an exchange economy*. Mimeo. Champaign-Urbana: University of Illinois.

Compensated Demand

Tatsuo Hatta

Hicks Compensation: Definition

When a consumer faces a price change under a given nominal income, his utility (or real income) level as well as his demand vector changes. Suppose, however, that his income level is

simultaneously changed as the price is changed so as to keep his utility at the initial level. This operation may be regarded as a compensation for the price change, and we call the resulting demand vector the *compensated demand* for the new price.

Thus the compensated demand is a function of the price vector and the utility level, and we may write it as

$$x = h(p, \mu), \tag{1}$$

where x and p are the consumption and price vectors, while μ is the utility level. We call h the *compensated* (*or Hicks*) *demand function*. Formally, it may be defined as the solution function of the following minimization problem:

$$\min_x p'x \text{ subject to } u(x) = \mu, \tag{2}$$

where u is the utility function.

Basic Properties

Hicksian Demand Rules

The Jacobian matrix of h with respect to p, denoted as h_p, is nothing but the Hicks substitution matrix. It has well-known properties:

$$p'h_p(p, \mu) = 0 \tag{3}$$

$$yh_p(p, \mu)y \leq 0 \text{ for all } y. \tag{4}$$

$$h_p(p, \mu) = h'_p(p, \mu). \tag{5}$$

Condition (3) is called the homogeneity condition, since it shows that the function h is homogeneous of degree zero with respect to p. Conditions (4) and (5) are called the negative semi-definiteness and the symmetry conditions, respectively. We will call these three conditions the *Hicksian Demand Rules*.

Shephard–Samuelson Lemma

The minimized expenditure value of problem (2) is a function of p and μ. This is called the *expenditure function*. Formally, we define it by

$$e(p, \mu) \equiv \min_x \{p'x | u(x) = \mu\}.$$

By definition, we obviously have

$$e_p(p, \mu) \equiv p'h(p, \mu). \tag{6}$$

There is a less obvious, but extremely useful, relationship between the compensated demand and the expenditure functions:

$$e_p(p, \mu) \equiv h(p, \mu). \tag{7}$$

This identity usually referred to as the *Shephard–Samuelson Lemma* was obtained by Hicks (1946, p. 331), Samuelson (1947, p. 68, 1953–54, pp. 15–16), and Shephard (1953).

To prove the Shephard–Samuelson Lemma, let x^* be an expenditure-minimizing vector that yields μ at the price p^*, i.e.,

$$x^* \equiv h(p^*, \mu). \tag{8}$$

Define the *gain function* g by

$$g(p) \equiv e(p, \mu) - p'x^*. \tag{9}$$

This and the definition of e imply that $g(p) \leq 0$. Also from (6) and (8), we have $g(p^*) = 0$. Hence the function g takes its minimum value of 0 when $p = p^*$. Therefore, the first and the second order minimization conditions yield

$$g_p(p^*) = 0 \tag{10}$$

and

$$y'g_{pp}(p^*)y \geq 0 \text{ for all } y \neq 0. \tag{11}$$

Equation 10 immediately proves (7).

To demonstrate the usefulness of the Shephard–Samuelson Lemma, let us prove the Hicksian Demand Rules from this Lemma. From (7) we have

$$e_{pp}(p, \mu) \equiv h_p(p, \mu).$$

This immediately yields (5). In view of (9) and (11), this also proves (4). On the other hand, (6), (9) and (10) yield

$$h(p^*, \mu) + p^{*'} h_p(p^*, \mu) - x^* = 0.$$

From (8), therefore, we obtain (3).

Uncompensated Demand

Reflection

The concept of the compensated demand is essential in analysing properties of the ordinary demand function. To see this, consider a maximization problem:

$$\max u(x) \quad \text{subject to} \quad y = p'x, \tag{12}$$

where y is the income level. Its solution function $m(p,y)$ is called the *ordinary* (or *Marshallian*) *demand function*. Define the *indirect utility function* by

$$v(p, y) \equiv u[m(p, y)]. \tag{13}$$

Problem (2) may be regarded as the expenditure minimization problem associated with (12). Newman (1982) calls it the *reflection* or the *mirror image* of maximization problem (12). If we let $\mu = v(p,y)$ in problem (2), the resulting minimum expenditure must equal y and the compensated demand must be equal to $m(p,y)$ in Problem (12). Thus we have

$$y \equiv e[p, v(p, y)] \tag{14}$$

and

$$m(p, y) \equiv h[p, v(p, y)]. \tag{15}$$

Roy's Identity

These identities yield *Roy's identity*,

$$v_p(p, y) \equiv - v_y(p, y) m(p, y). \tag{16}$$

To see this, differentiate (14) with respect to p and y to get

$$e_\mu[p, v(p, y)] v_p(p, y) \equiv - e_p[p, v(p, y)] \tag{17}$$

and

$$1 \equiv e_\mu[p, v(p, y)] v_y(p, y), \tag{18}$$

respectively. Multiplying by $v_y(p,y)$ on both sides of (17), and then applying (18), (7) and (15), we obtain (16).

Slutsky–Hicks Decomposition

Identities (14) and (15) also yield the *Slutsky–Hicks decomposition*,

$$m_p(p, y) \equiv h_p[p, v(p, y)] - m_y(p, y) m(p, y). \tag{19}$$

Thus the slope of the ordinary demand function equals the slope of the compensated demand function adjusted to the income effect.

To prove (19), differentiate (15) with respect to p and y to get

$$m_p(p, y) \equiv h_p[p, v(p, y)] + h_\mu[p, v(p, y)] v_p(p, y) \tag{20}$$

and

$$m_y(p, y) \equiv h_\mu[p, v(p, y)] v_y(p, y), \tag{21}$$

respectively. The only difference between (19) and (20) is their income terms. Applying (16) and (21) to the last term of (20), we get (19).

Slutsky Compensation

After a price change takes place, the Hicks compensation keeps the consumer on the same utility level as before the price change. As Mosak (1942) pointed out, however, Slutsky had a different concept of compensating the loss of real income. Slutsky considered a compensation that 'makes possible the purchase of the same quantities of all the goods that had formerly been bought', When a price change takes place, the Hicks-compensated and the Slutsky-compensated demand effects are generally different. When the price change is infinitesimal, however, they become equal, and this equality is called Mosak's Equality.

Mosak's equality has played an important role in index number theory. The Laspeyres index, which is widely adopted in practice, is based on the Slutsky compensation, since it indicates the change in income that would be needed in the current year in order to buy the commodity bundle bought in the base year. Under the Slutsky substitution effect, the individual can be no worse off and is likely to be better off since he is able to purchase at least the bundle he had before the price change. Thus he is 'overcompensated' for the price change (see Samuelson 1953, pp. 4–5). The price index that truly reflects the utility change should be based on the Hicks compensation. Such an index is difficult to compute because utility levels are not observable. Mosak's equality reveals, however, that for small price changes the Laspeyres index is a good approximation to the 'ideal' index.

Let us now formally state Mosak's equality. Define the function s by

$$s(p,x) \equiv m(p,p'x). \tag{22}$$

The function s is the demand function with the fixed endowment bundle x. We call $s_p\,[p, m\,(p, y)]$ the *Slutsky substitution matrix*. It represents the variation in demand when the price change is accompanied by an income compensation that keeps the original consumption bundle $m(p,y)$ on the budget plane. Mosak's Equality may now be expressed as

$$s_p[p, m(p,y)] \equiv h_p[p, v(p,y)]. \tag{23}$$

To prove this, first define the function w by

$$w(p,x) \equiv v(p,p'x). \tag{24}$$

The value of $w(p,x)$ represents the maximized utility level when the endowment bundle x is given. Differentiating this and applying Roy's identity, we get

$$w_p[p, m(p,y)] = 0. \tag{25}$$

Thus, if the utility maximizing bundle under a certain price vector happens to be equal to the endowment bundle, the utility level is hardly changed by a slight change of the price vector away from the initial one. Equations (15), (22) and (24), yield

$$s(p,x) = h[p, w(p,x)].$$

Differentiating this with respect to p and noting (24), we immediately have (23).

Historical Notes

Slutsky (1915) first established the homogeneity and symmetry conditions on the substitution matrix. Since he did not have the concept of utility-maintaining compensation or the compensated demand function, he derived these properties for the Slutsky compensated substitution matrix $s_p\,[p, m(p, y)]$ rather than for the Hicksian matrix $h_p\,[p, v(p, y)]$. During the 1930s, Hicks and Allen (1934) and Hicks (1939, 1946) gave verbal interpretations to the substitution matrix in terms of the Hicksian compensation. But in their formal derivation of its properties, they defined the substitution matrix to be the Slutsky substitution matrix, as is clear from the following passage from the Mathematical Appendix of Hicks (1946, p. 309):

> . . .it follows from the equation that the substitution term represents the effect on the demand for xs of a change in the price of xr combined with such a change in income as would enable the consumer, if he chose, to buy the same quantities of all goods as before, in spite of the change in Pr.

Thus they too did not explicitly state the function h, much less gave a name to it.

To the writer's knowledge, Samuelson (1947) is the first author who explicitly stated (1) and derived Hicksian rules directly from it, though he did not give a name to it (see Samuelson 1947, (43) on p. 103 and (99) on p. 114). Subsequently, Samuelson (1953, p. 8, n1) gave a heuristic proof of the symmetry and the negative-semidefiniteness condition of the Slutskycompensated substitution matrix as envelope properties, i.e. in a spirit very much similar to the one given above. In this path-breaking proof, however, he relied upon the indirect utility

function rather than the expenditure function, thus without using the mirror image minimization problem or the compensated demand function.

McKenzie (1956) and Karlin (1959) explicitly defined the function *h*, and derived its properties by taking full advantage of the Shephard– Samuelson Lemma. However, they had to use a *global* property of the expenditure function in deriving the negative semidefiniteness condition, which is a *local* property. The proof of this condition employed above is solely based on a *local* minimization condition, and serves as the mirror image counterpart of Samuelson's (1953) proof for the Slutsky substitution matrix. Diamond and McFadden (1973) attribute this proof method to Gorman (see Gorman 1976). The above proof of Mosak's equality is due to Hatta and Willke (1982).

Silberberg (1974) considered the general maximization problem with possibly many constraint functions where both target and constraint functions may be non-linear. Extending Samuelson's (1965) proof method, he showed that the compensated solution function of that problem satisfies generalized forms of the symmetry and non-negative definite conditions, as long as the constraint functions do not contain shift parameters. Silberberg's proof boils down to Gorman's in the standard expenditure minimization problem. Hatta (1980) extended the concept of the compensated demand function to the case where the same shift parameters may appear simultaneously in both target and constraint functions in the general problem. He showed that the properly compensated solution function in that problem satisfies generalized forms of symmetry and non-negative definite conditions. He also established an envelope theorem that contains both the Shephard–Samuelson Lemma and Roy's Identity as special cases. His proof integrates Samuelson's (1953) and Gorman's into one. The global characterization of the compensated demand function by McKenzie and Karlin was extended into *duality theory*, as surveyed by Diewert (1982).

In many branches of economics outside the demand theory, the concept of the compensated demand function was implicitly used without being explicitly stated. Examples are Hotelling's (1938) and Harberger's (1974) analysis of the *excess burden of taxation*, Hicks's (1956) *compensating* and *equivalent variation* that illuminate the concept of consumers' surplus, and Alonso's (1964) *rent bid function*, which keeps a consumer's utility constant regardless of the location he chooses. A number of economists of the Chicago School, including Friedman (1949), Bailey (1954), and Becker (1971), used the concept of the Hicks compensation in various welfare analyses. Each of these authors gave a different name to the concept. Friedman called its graph the *Marshallian demand curve* contrary to the current usage of this term; Bailey, the *constant-real-income demand curve*, and Becker, the *pure demand curve*.

The explicit use of the compensated demand function gave rise to dramatically clearer restatements and proofs of many existing theorems. Its usefulness has reached far beyond that, however. Since the early 1970s, this function has been used for the analyses of the welfare impacts of parametic shifts in various general equilibrium models, as stated by Ohyama (1974), Takayama (1974), Diamond and MacFadden (1974), Dixit (1975) and Hatta (1977, 1980), and comprehensively studied by Dixit and Norman (1980) and Woodland (1982).

The history of the compensated demand function is curious. The properties of its derivatives were known and the concept of Hicks compensation used in many fields of economics before the function itself was stated or named. Perhaps this is because economists has an unconscious reluctance in putting the elusive concept of the utility level as a variable of a function. Once explicitly stated and well understood, however, the compensated demand function has found a powerful use in welfare economics, precisely because it has the utility level, rather than income, as an explicit variable.

See Also

- ▶ Demand Theory
- ▶ Duality
- ▶ Index Numbers

Bibliography

Alonso, W. 1964. *Location and land use*. Cambridge, MA: Harvard University Press.

Bailey, M.J. 1954. The Marshallian demand curve. *Journal of Political Economy* 62: 255–261.

Becker, G.S. 1971. *Economic theory*. New York: Knopf.

Diamond, P.A., and D.L. McFadden. 1974. Some uses of expenditure function in public finance. *Journal of Public Economics* 3: 3–21.

Diewert, W.E. 1982. Duality approaches to microeconomic theory. In *Handbook of mathematical economics*, vol. 2, ed. K.J. Arrow and M.D. Intriligator, 535–599. Amsterdam: North-Holland.

Dixit, A.K. 1975. Welfare effects of tax and price changes. *Journal of Public Economics* 4: 103–123.

Dixit, A.K., and V. Norman. 1980. *Theory of international trade*. Cambridge: Cambridge University Press.

Friedman, M. 1949. The Marshallian demand curve. *Journal of Political Economy* 57: 463–495.

Gorman, W.M. 1976. Tricks with utility functions. In *Essays in economic analysis*, ed. M. Artis and R. Nobay. Cambridge: Cambridge University Press.

Harberger, A.C. 1974. *Taxation and welfare*. Boston: Little, Brown.

Hatta, T. 1977. A recommendation for a better tariff structure. *Econometrica* 45: 1859–1869.

Hatta, T. 1980. Structure of the correspondence principle at an extremum point. *Review of Economic Studies* 47: 987–997.

Hatta, T., and R.J. Willke. 1982. Mosak's equality and the theory of duality. *International Economic Review* 22: 361–364.

Hicks, J.R. 1939. *Value and capital*. London: Oxford University Press.

Hicks, J.R. 1956. *A revision of demand theory*. London: Oxford University Press.

Hicks, J.R., and Allen, R.D.G. 1934. A reconsideration of the theory of value, I, II. *Econometrica*, N.S.I, 52–75, 196–219.

Hotelling, H.S. 1938. The general welfare in relation to the problems of taxation and of railway and utility rates. *Econometrica* 6: 242–269.

Karlin, S. 1959. *Mathematical methods and theory in games programming and economics*, vol. 1. Reading: Addison-Wesley Publishing Company.

McKenzie, L. 1956. Demand theory without a utility index. *Review of Economic Studies* 24: 185–189.

Mosak, J. 1942. On the interpretation of the fundamental equation of value theory. In *Studies in mathematical economics and econometrics*, ed. O. Lange, 69–74. Chicago: University of Chicago Press.

Newman, P.K. 1982. Mirrored pairs of optimization problems. *Economica* 49: 109–119.

Ohyama, M. 1974. Tariffs and the transfer problem. *Keio Economic Studies* 11(1): 29.

Samuelson, P.A. 1947. *Foundations of economic analysis*. Cambridge, MA: Harvard University Press.

Samuelson, P.A. 1953. Consumption theorems in terms of overcompensation rather than indifference comparison. *Economica* 20(February): 1–9.

Samuelson, P.A. 1965. Using full duality to show that simultaneously additive direct and indirect utilities implies unitary price elasticity of demand. *Econometrica* 33: 781–796.

Shephard, R. 1953, 1970. *Cost and production functions*. Princeton: Princeton University Press.

Silberberg, E. 1974. A revision of comparative statics methodology in economics, or, how to do comparative statics on the back of an envelope. *Journal of Economic Theory* 7: 159–172.

Slutsky, E. 1915. Sulla teoria del bilancio del consumatore. *Giornale degli Economisti* 51: 1–26. English trans. in *Readings in price theory*, ed. G.J. Stigler and K.E. Boulding, Chicago: Chicago University Press, 1953.

Takayama, A. 1974. On the analytical framework of tariffs and trade policy. In *Trade stability and macroeconomics, essays in honor of Lloyd A. Metzler*, ed. G. Horwich and P.A. Samuelson, 153–178. New York: Academic.

Woodland, A.D. 1982. *International trade and resource allocation*. Amsterdam: North-Holland.

Compensating Differentials

Matthew E. Kahn

Abstract

Compensating differentials represent a wage premium for unpleasant aspects of a job. Jobs differ along several dimensions. Some jobs offer generous health insurance benefits. Others entail long hours or may expose workers to physical risks. Some are available only in polluted cities. In equilibrium, labour markets accommodate diversity by establishing wages that tend to make different jobs relatively close substitutes at the margin. Using hedonic wage regression techniques, researchers have estimated the equilibrium implicit market price that workers pay, through lower wages, for working in a more pleasant setting. This technique is widely used by labour and environmental economists.

Keywords

Compensating differentials; Hedonic wage function; Labour economics; Rosen, S.; Superstars, economics of; Unobserved skill; Urban economics; Urban environment and quality of life; Wage heterogeneity, sources of; Wage premium; Worker heterogeneity

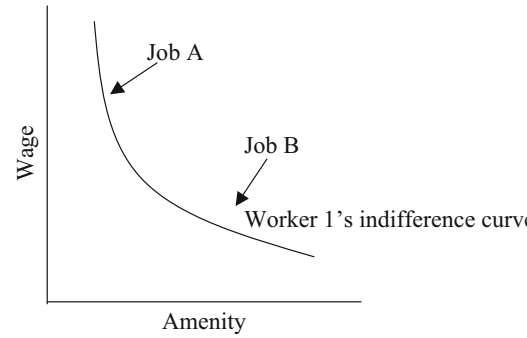
JEL Classifications

J300

Compensating differentials represent a wage premium for unpleasant aspects of a job. Jobs differ along a number of dimensions. Some jobs offer generous health insurance benefits. Other jobs entail long hours or may expose workers to physical risks. Some jobs are only available in polluted cities. The theory of compensating differentials is based on the simple premise that there is 'no free lunch'. In market equilibrium, more unpleasant jobs will offer a wage premium relative to other jobs. Similarly, homes in nicer communities or high-quality-of-life cities will sell for a premium. To quote Sherwin Rosen (2002, p. 2), 'Markets accommodate diversity by establishing prices that tend to make different things relatively close substitutes at the margin. Adam Smith's insight that market prices tend to equalize their net advantages is fundamental to these problems. If one good has more desirable characteristics than another, the less preferred variety must compensate for its disadvantages by selling at a lower price.'

Defining Compensating Differentials

Jobs represent tied bundles of attributes. Suppose that a worker gains utility from earning a wage and from a job attribute. This attribute could represent job safety, or total days of vacation, or health insurance benefits. As shown in Fig. 1, there are two jobs, A and B. Each job represents a different bundle of a wage and a non-market job-specific amenity level. The two jobs differ: job B is the more pleasant of the two. If all workers have the same utility function, then in equilibrium this representative worker must be indifferent between the two jobs. Thus, job A must pay a higher wage than job B to compensate this worker.

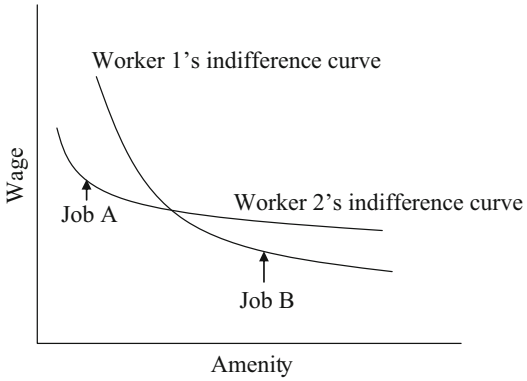


Compensating Differentials, Fig. 1

The econometrician can collect data on each job type's wage and amenities. In a more realistic economy where there are many types of jobs that differ with respect to the wage and their amenity level, the representative worker's indifference curve would be sketched out. The slope of the representative worker's indifference curve represents the compensating differential of how much lower a wage this worker would accept in return for a small increase in the job amenity.

To see how worker heterogeneity affects the interpretation about observed compensating differentials, consider the simple extension where we introduce two types of workers. These workers are equally productive but differ with respect to their demand for working in the more pleasant job. In Fig. 2, worker 1 values the job amenity more than worker 2. In equilibrium, job A will pay a compensating differential to attract workers to be willing to work in this job. Worker 2 will choose to work in job A while worker 1 will choose to work in job B. Firm A will prefer to hire worker 2 rather than worker 1 because worker 1 requires a larger compensating differential for working in the more unpleasant job. The profit maximizing firm seeks to minimize its costs of production.

The econometrician will observe the equilibrium wage paid to workers in job A and B. As shown in Fig. 2, this equilibrium wage–amenity relationship called the *hedonic wage function*



Compensating Differentials, Fig. 2

does not represent either worker 1's or worker 2's indifference curves. Instead, this hedonic wage function represents the envelope of the minimum wage that heterogeneous workers are willing to accept to do a job. This simple example highlights how introducing worker heterogeneity affects inference from observed data (see Rosen 2002). Figure 2 focuses on just one dimension of worker heterogeneity. The recent compensating differentials literature has explored the consequences of other dimensions of worker heterogeneity such as unobserved skill (IQ, for example) and a worker's ability to self-protect against injury on the job (Hwang, Reed and Hubbard 1992; Shogren and Stamland 2002).

Labour Econometric Applications of Compensating Differentials Theory

An enormous applied econometrics literature has estimated various versions of hedonic wage functions to recover estimates of the marginal valuation of non-market job attributes. One major focus of this research has been to estimate the value of life by measuring how much of a wage premium the marginal worker requires for working in a job with a higher probability of death (Viscusi and Aldy 2003). Other studies have used hedonic methods to measure the compensating differential for mandated government health insurance benefits (Gruber 1994).

The standard approach utilizes a large micro-data set. The dependent variable in such a study is a full-time worker's wage in a specific occupation, industry or city. For example, in Eq. 1 the dependent variable is the log of worker i's wage in industry j in year t. In an urban application, j would refer to a city rather than an industry. The researcher will include a large number of demographic controls, such as age, ethnicity, or education, to 'standardize' the worker. If one controls for these factors, the key variables of interest are the Z's in Eq. 1. In a labour economics application, the Z vector may represent a set of job specific attributes (length of day, job risk). In an urban economics application, the Z vector may represent attributes of the city where the job is located (climate, pollution, crime).

$$\mathrm{Log}\left(\mathrm{Wage}_{ijt}\right) = \gamma_0 + \gamma_1^* X_{it} + \gamma_2^* Z_{jt} + U_{ijt} \quad (1)$$

Ordinary least squares regression estimates of γ_2 are used to construct measures of the compensating differentials for job tasks and characteristics of employment locations. Estimates of such coefficients have been used to rank city quality of life (see Gyourko and Tracy 1991) and represent the first stage of the hedonic two-step for recovering demand functions for non-market goods such as air quality or climate (Rosen 1974; Ekeland et al. 2004).

If the population differs with respect to its tastes for job attributes, then γ_2 can be used to construct a worker's budget constraint. For example, in a job-safety regression if γ_2 equals minus \$100 then this means that a one-unit increase in job safety will cost the worker an extra \$100 in wages. The rational worker facing this budget constraint will take this trade-off into account when choosing the job that maximizes her utility.

Hedonic estimates of compensating differentials can also be used to bound worker preferences. To return to Fig. 2, a lower bound on worker 1's willingness to accept work in risky job A is the equilibrium wage paid to worker 2. Since we know that worker 1 chose the safe job and refused to work in job A at the wage that worker 2 accepted, worker 2's wage offer provides a lower bound (see Rosen 2002).

The typical hedonic wage regression study estimates Eq. 1 using ordinary least squares. This econometric approach will yield consistent estimates of y_2 if the unobserved determinants of wages (that is, the error term) are uncorrelated with the explanatory variables. What is the error term in this hedonic pricing equation? While a researcher might hope that it represents measurement error in the dependent variable, it is more likely that the error term represents unobserved attributes of the worker and unobserved attributes of the geographical area where the worker lives and works.

Unfortunately for researchers, people self-select where to live and work. A researcher would like to know what wage the same worker would earn in every industry and in every city. In a cosmopolitan city such as New York, superstars of all fields, ranging from Don Trump in real estate to Derek Jeter in baseball, have all chosen to work there. A naive cross-city hedonic researcher would observe these stars living in New York City earning high wages *relative* to observationally identical people in Tulsa, and would conclude, based on the wage regression, that New York City's quality of life must be worse than Tulsa's. Clearly, the problem with this inference is the 'apples to oranges' comparison. New York City's amenities are a normal good. The high-skilled earn higher salaries and are attracted to living and working in this city.

Conclusion

A job's wage is not a sufficient statistic for its quality. Coal miners are paid a relatively high wage but the work is dangerous and unpleasant. A major research agenda in labour economics investigates how much people implicitly pay for non-market job attributes. Credible estimates of wage compensating differentials for living in less polluted cities or working in risky industries would greatly aid policy analysis that seeks to measure the benefits of environmental and safety regulation.

See also

- ▶ Roy Model
- ▶ Wage Inequality, Changes in

Bibliography

Ekeland, I., J. Heckman, and L. Nesheim. 2004. Identification and estimation of hedonic models. *Journal of Political Economy* 112: S60–S109.

Gruber, J. 1994. The incidence of mandated maternity benefits. *American Economic Review* 84: 622–641.

Gyourko, J., and J. Tracy. 1991. The structure of local public finance and the quality of life. *Journal of Political Economy* 91: 774–806.

Hwang, H.-S., R. Reed, and C. Hubbard. 1992. Compensating wage differentials and unobserved productivity. *Journal of Political Economy* 100: 835–858.

Rosen, S. 1974. Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy* 82: 34–55.

Rosen, S. 2002. Markets and diversity. *American Economic Review* 92: 1–15.

Shogren, J., and T. Stamland. 2002. Skill and the value of life. *Journal of Political Economy* 110: 1168–1173.

Viscusi, W., and J. Aldy. 2003. The value of statistical life: A critical review of market estimates throughout the world. *Journal of Risk and Uncertainty* 27: 5–76.

Compensation Principle

John S. Chipman

Abstract

The compensation principle holds that one of two possible states constitutes an improvement over the other if the gainers could compensate the losers for their losses and still be at least as well off as in the original state. The conflict between potentiality and actuality – one situation is judged better than another if everybody *could* be made better off in the new situation even though some in fact become worse off – ensures that the compensation principle does not allow for value-free policy decisions.

Keywords

Autarky; Baldwin envelope; Barone, E.; Bergson–Samuelson social-welfare function; Bickerdike, C. F.; Cairnes, J. E.; Compensation principle; Competitive equilibrium; Consumer surplus; Cost–benefit analysis; Deadweight loss; Debreu, G.; Dupuit, A.-J.-E. J.; Exchange; Excise taxes; Free trade; Fundamental theorem of welfare economics; Gains from trade; Game theory; General equilibrium; Hicks, J. R.; Hotelling, H.; Ideal money; Income-compensation function; Interpersonal utility comparisons; Kaldor, N.; Kuznets, S.; Lerner, A. P.; Little, I. M. D.; Lump sum taxes; Marshall, A.; Mishan, E. J.; Monopoly; New welfare economics; Optimal tariffs; Pantaleoni, M.; Pareto optimality; Partial equilibrium; Pigou, A. C.; Proportional income tax; Revealed preference; Ricardo, D.; Robbins, L. C.; Samuelson, P. A.; Scitovsky indifference surface; Scitovsky, T.; Taxation of income; Transferable utility; Utility-possibility frontier; Value judgements; Viner, J.

JEL Classifications

B4

The term 'compensation principle' refers to the principle that, in comparing two alternative states in which a given community of persons might find itself, one of the states constitutes an improvement over the other (in the weak sense including equivalence) if it is possible for the gainers to compensate the losers for their losses and still be at least as well off as in the original state.

If the hypothetical compensation is actually carried out, the principle reduces to the Pareto criterion: all are at least as well off, in one state compared to the other. There is no need to invoke the compensation principle in such a case. On the other hand, if the principle is used to compare two unique alternative states in which a community might find itself, neither of which is Pareto-superior to the other, the principle seems quite arbitrary unless interpreted in a broader context. There is a sense in which one person might be said to be basically healthier than another even though, at the particular moment, such a person might have a cold and the other one not. The compensation principle is usually used to make comparisons in this sense; one state of the economy is sounder, healthier, more robust, or has greater productive potential, than another. What this implies is that states under comparison are usually not unique, singleton states but composite ones, or sets of states. Formally, the objects being compared are usually sets of commodity bundles that could be made available to the aggregate of consumers, described in the literature as 'situations' in contrast to single 'points' in such sets (cf. Baldwin 1954).

Examples of comparisons in which the compensation principle is typically used are those between (*a*) a perfectly competitive system of industrial organization and an imperfectly competitive one; (*b*) free trade and no trade (or restricted trade); (*c*) the state of an economy before and after a war, or depression, or change in productive techniques. Most but not all of these types of comparisons are relevant to policy decisions; and the policy decisions are usually not of an ad hoc type (for which the compensation principle would hardly be appropriate) but of a fundamental nature concerning the underlying system of industrial organization and trade.

Inasmuch as the principle can be applied without the need to make interpersonal comparisons, some of its more ardent proponents have maintained that it is 'value-free'. However, there can be no doubt that it does require acceptance of some value judgements, since the Pareto criterion itself constitutes one – albeit a widely accepted one. Another value judgement implicit in the principle as it has usually been applied is that each individual is the best judge of his or her own well-being; while also quite widely accepted, this one is obviously controversial, and in fact government policy measures are often called for precisely in those instances where it is clearly an untenable assumption. But the most important and controversial way in which value judgements enter into the compensation principle is in the conflict between potentiality and actuality: one situation is judged better than another if everybody *could* be made better off in the new situation even though some in fact become worse off. This

lacuna in the principle has led Little (1950) and Mishan (1969) to formulations in which compensation tests are combined with explicit distributional value judgements, and Samuelson (1947, 1956) into a full-fledged ethical system in which compensation is carried out to the extent that the ethical norms dictate.

In many applications the compensation principle is difficult to formulate in a precise manner unless one assumes absence of externalities in consumption, so it is usually formulated (but with some notable exceptions – for example, Coase 1960) under the assumption that each person's welfare depends only on his or her own consumption of goods and services. In most applications, the data available for making comparisons are, almost inevitably, limited to aggregative information on the actual state of the economy in each situation; much of the work in applying the principle therefore consists in using economic theory to make inferences from the actual observations concerning underlying conditions in the economy. By its nature, the compensation principle is limited in its application to comparing alternative states (or sets of states) of a given community of individuals; thus, it cannot be applied (at least not literally) to historical comparisons of a country's condition over time (since the population has changed) or to comparisons of the living conditions of different countries (since the populations are different). However, extensions of the principle to cover such comparisons are possible provided suitable additional empirical assumptions and value judgements are accepted; for example, if all individuals are assumed to have identical preferences, one could ask whether there exists a redistribution of income in each period (or country) such that each individual in the one situation would be better off than each individual in the other. This would obviously entail additional value judgements along with the additional empirical assumptions.

Historical Development: From Dupuit to Hotelling

The compensation principle may be traced back to Dupuit (1844, pp. 359–60; Arrow and Scitovsky 1969, p. 272) and Marshall (1890, p. 447; 1920, p. 467) who used the concept of consumers' surplus to compare the losses of consumers (say from a bridge toll or an excise tax) with the gains to the government. The demonstration that the former exceed the latter, so that consumers cannot be compensated for their losses out of the government revenues, provided a convincing case for the superiority of income tax to an excise tax (or for the superiority of government subsidization of bridge construction to its financing of it by tolls), and at the same time provided scientific prestige and great intuitive appeal to a method that was able to reach such a definitive conclusion and furnish a measure of the 'deadweight loss'.

While Dupuit and Marshall used partial-equilibrium analysis, Pareto (1894, p. 58) was the first to introduce the concept into general-equilibrium theory, in the course of an article devoted to proving the optimality of competitive equilibrium. In the first part of this article (summarized by Sanger 1895), Pareto used as his criterion of optimality the sum of individual utilities; in the second part, however – acknowledging the criticisms and suggestions of Pantaleoni and Barone (both admirers of Marshall, which Pareto was not) – he reformulated the problem so as to sum not the utilities of different consumers but the quantities they consume. His criterion of optimality (1894, p. 60) was that it should be impossible for one person to gain without another losing – 'Pareto optimality' – a criterion that had also been introduced by Marshall (1890, pp. 449–50; 1920, pp. 470–1). A more refined version of Pareto's argument later appeared in the *Cours* (Pareto 1896–7, vol. 1, pp. 256–62; vol. 2, pp. 88–94).

The proposition formulated by Pareto (1894) anticipated what has now come to be known as the 'fundamental theorem of welfare economics', namely, that every competitive equilibrium is Pareto optimal and, conversely, every Pareto optimum can be sustained by a competitive equilibrium. Pareto considered the problem faced by a socialist state striving to attain an outcome in which it was impossible for one person to gain without another losing. The Ministry of Justice would concern itself with problems of income

distribution, and the Ministry of Production with resource allocation and choice of production coefficients. A weakness of Pareto's argument was that he assumed a price system already to be established – perhaps our socialist state needs the prices of its capitalist neighbours to guide it. Pareto further assumed that each individual's budget constraint was adjusted by the addition of a parameter (a lump-sum subsidy or tax) controlled by the government. The government's objective was to maximize the sum of these parameters, which he showed was equal to aggregate profit – the value of commodities consumed less the value of factor services supplied, equal to the value of firms' output less the outlay on their factor inputs. If it were possible to increase all the parameters, the existing situation would not be Pareto optimal; if their sum were a maximum, it would not be possible to increase one of them without decreasing another, and the outcome would be Pareto optimal. Pareto showed that maximization of aggregate profit at the given prices, subject to the resource-allocation and production-function constraints, would lead to cost-minimization and zero profits. (For mathematical details of Pareto's arguments see Chipman 1976, pp. 88–92). Pareto summarized this result by stating (1896–7, vol. 2, p. 94):

> Free competition of entrepreneurs yields the same values for the production coefficients as would be obtained by determining them by the condition that commodity outputs should be chosen in such a way that, for some appropriate distribution, maximum ophelimity would be achieved for each individual in society.

The last clause was Pareto's unfortunately awkward way of stating the criterion of Pareto optimality.

Barone (1908), who had originally spurred Pareto on to this line of argument, developed it further himself. He noted that a competitive equilibrium has the property that aggregate profit is at a maximum at the equilibrium prices, hence, for any feasible departure from this equilibrium, valuing consumption and factor services at the equilibrium prices, some individuals may gain and others will lose, the losses outweighing the gains so that, even if the gainers part with all their gains, the rest will still be worse off than originally. (Barone used what is now known as the criterion of revealed preference to make inferences concerning preferences from data on prices and incomes). Such a state was described by Pareto and Barone as 'destruction of wealth', and its measure by aggregate income loss at the competitive-equilibrium prices provided an alternative to the deadweight loss considered by Dupuit and Marshall. Barone (1908) also related his arguments to those of Marshallian consumers'-surplus analysis.

Lerner (1934) invoked the compensation principle in his proposed method for measuring monopoly power, describing it as 'a loss to the consumer which is not balanced by any gain reaped by the monopolist'. In this paper Lerner also formulated, apparently independently, the concept of Pareto optimality.

Hotelling (1938) made a noteworthy contribution by providing an alternative demonstration of the inferiority of excise taxes to income taxes, using the compensation principle directly. He considered a single individual consuming n commodities in amounts q_j and facing market prices p_j. Prior to the imposition of the excise taxes (or tolls), the individual consumes a bundle q^0 at prices p^0 and income (or fixed component of income) m^0, which maximizes a utility function $U(q)$ subject to the budget constraint $p^0 \cdot q = m^0$. Subsequent to the introduction of taxes, market (tax-inclusive) prices and after-tax income are p^1 and m^1 respectively, and a bundle q^1 is chosen which maximizes $U(q)$ subject to $p^1 \cdot q = m^1$. The government collects $r = (p^1 - p^0) \cdot \mathrm{q}^1 - (m^1 - m^0)$ in revenues. Since the government is assumed to collect $\left(p_j^1 - p_j^0\right) \cdot q_j^1$ in taxes on commodity j,p_j^0 must be identified with the production cost after the tax (as well as with the market price = production cost before the tax); this is a fairly restrictive assumption, since it implies that the tax does not affect production costs. (In this respect Hotelling's treatment is less general than Dupuit's and Marshall's, involving infinite elasticities of supply). We may denote the ad valorem excise-tax rate on commodity j by $t_j = p_j^1/p_j^0 - 1$, and a proportional income-tax rate by $t_0 = 1 - m^1/\mathrm{m}^0$ (negative

taxes are interpreted as subsidies). The government's revenues are

$$r = \sum_{j=1}^{n} t_j p_j^0 q_j^1 + t_0 m^0 = 0,$$

assumed zero since the government distributes the total proceeds of these excise taxes back to the consumer (or taxes the consumer if these are negative). The consumer's budget constraint after the imposition of the taxes is

$$\sum_{j=1}^{n} (1 + t_j) p_j^0 q_j^1 = (1 - t_0) m^0.$$

These two equations together imply that q^1 satisfies the budget constraint $p^0 \cdot q^1 = m^0$, hence q^1 was in the consumer's original budget set. Therefore, setting aside the 'infinitely improbable ... contingency' that q^0 and q^1 lie on the same indifference surface, Hotelling concluded (1938, p. 252) that 'if a person must pay a certain sum of money in taxes, his satisfaction will be greater if the levy is made directly on him as a fixed amount than if it is made through a system of excise taxes which he can to some extent avoid by rearranging his production and consumption'.

Unfortunately Hotelling overlooked the fact that if $t_j = t$ for all j then the government's budget constraint implies $p^0 \cdot q^1 = -m^0 t_0/t$, whence $t_0 = -t$ and $q^1 = q^0$. That is, a system of uniform ad valorem excise taxes is equivalent to a proportional income tax. This was pointed out by Frisch (1939) and accepted by Hotelling (1939). As Frisch made clear, what Hotelling really proved was the non-optimality of a system of non-proportional excise taxes or subsidies when selling prices are given. If these selling prices are equal to marginal costs, Hotelling's theorem shows that market prices should be proportional to marginal costs. Since incomes are fixed in Hotelling's formulation, income taxes may be regarded as lump-sum taxes. If institutional consideration make excise taxes impossible for one commodity (say leisure), then they must be zero for all commodities and optimality requires that prices be equal to marginal costs. (For a less charitable interpretation of Hotelling's contribution see Silberberg 1980).

Hotelling went on to assert that his proposition could be extended to many consumers (though no details or proof were provided), and he proceeded to examine the consumers'-surplus measure of loss $\frac{1}{2}(p^1 - p^0) \cdot (q^1 - q^0) = \frac{1}{2} T p^0 \cdot (q^1 - q^0)$ (where T is a diagonal matrix of excise-tax rates t_j). He also made some general observations (1938, p. 267) that, to this day, constitute what is probably the best statement to be found of the philosophy underlying the compensation principle.

The Years of the New Welfare Economics

In the cases to which the compensation principle was applied by Dupuit, Marshall, Lerner and Hotelling, compensation was made between the class of consumers on the one hand and a government or a monopolist on the other. While Pareto and Barone had discussed compensation between different classes of consumers (as had Hotelling in his general remarks) their work was unknown to English-speaking economists until the publication in 1935 of the English translation of Barone's 1908 work. Even this seems not to have struck home, however, since Kaldor (1939) cited passages from Harrod (1938) and Robbins (1938) to the effect that, since movement towards free trade would affect different classes differently, no scientific statement could be made concerning the beneficial effect of free trade without making interpersonal comparisons of utility.

Kaldor (1939) proceeded to sketch an argument to the effect that removal of an import duty (using the classical example of repeal of the Corn Laws) would result in a situation in which the losses incurred by the landlords could be compensated by the gains (through lower import prices) obtained by the other consumers. Such an argument cannot be correct, however, since, as Kaldor (1940) pointed out only a year later, it follows from Bickerdike's theory of optimal tariffs that a country can gain from the imposition of a sufficiently small duty, and, as Graaff (1949) and others later demonstrated, the compensation principle can be used to show that, with suitable

compensation, all persons can gain. Unless the rate of corn duty was above the optimal tariff rate, the opposite conclusion would follow to that indicated by Kaldor (1939).

A previous attempt by Pareto (1895) to show by means of the compensation principle that a tariff would lead to 'destruction of wealth' was defective, since he assumed trade to be balanced in domestic prices and thus he failed to take account of the improvement in the terms of trade and the beneficial effect of the tariff revenues.

Other attempts prior to 1939 to make the case for free trade suffered from vagueness both in specifying the criterion of gain and in specifying the alternative with which free trade was being compared. Ricardo (1815, p. 25) stated: 'There are two ways in which a country may be benefited by trade – one by increase of the general rate of profits . . . the other by the abundance of commodities, and by a fall in their exchangeable value, in which the whole community participate'. According to Cairnes (1874, p. 418), 'the true criterion of the gain on foreign trade [is] the degree in which it cheapens commodities, and renders them more abundant'. A hint of a compensation principle is found in Viner (1937, pp. 533–4):

> free trade . . . necessarily makes *available* to the community *as a whole* a greater physical real income in the form of more of *all* commodities, and . . . the state . . . can, by appropriate supplementary legislation, make certain that removal of duties shall result in more of *every* commodity for *every* class of the community.

Like Kaldor's statement, this is formally incorrect; but it was sufficiently suggestive to stimulate Samuelson (1939) into providing a formal proof of a gains-from-trade theorem, albeit under very restrictive assumptions.

Samuelson (1939) assumed that an open economy had a locus $\varphi(y, l) = 0$ of efficient combinations of outputs y and (variable) factor services l, and asserted that vectors of prices p and factor rentals w in competitive equilibrium would be such that aggregate profit $p \cdot y - w \cdot l$ is a maximum. This is the same as the proposition of Pareto (1894), and Barone (1908) referred to above. Letting x denote the bundle of commodities consumed, under both (balanced) free trade and autarky the budget equation $p \cdot x = p \cdot y$ holds. Letting superscripts 0 and 1 denote equilibrium values under autarky and free trade respectively, it follows that

$$p^1 \cdot x^1 - w^1 \cdot l^1 \geq p^1 \cdot x^0 - w^1 \cdot l^0.$$

Assuming all N individuals to be identical in their preferences and ownership of factors, and dividing this inequality through by the number of individuals, it states that each person chooses $(x^1/N, l^1/N)$ under free trade when $(x^0/N, l^0/N)$ is available, hence (if $p^1 \neq p^0$) each person prefers $(x^1/N, l^1/N)$ to $(x^0/N, l^0/N)$. Therefore free trade is Pareto-superior to autarky.

Samuelson went on to assert (1939, p. 204) that, if the assumption of identical individuals is dropped, then, although it could no longer be said that each individual was better off under free trade, 'it would always be possible for those who desired trade to buy off those opposed to trade, with the result that all could be made better off'. This argument went unchallenged until Olsen (1958) pointed out that, if compensation were paid from gainers to losers, a new equilibrium price constellation p^1 would result, and the argument no longer follows. For this reason Samuelson's 1939 results has come to be known as the gains-from-trade theorem for the 'small-country case', though this interpretation was not suggested by Samuelson at the time. But this description of Samuelson's result is inaccurate. Generalizing his argument we can say that if (x_i^t, l_i^t) are the allocations of (x^t, l^t) to individual i, where $\sum_{i=1}^{N} x_i^t = x^t$ and $\sum_{i=1}^{N} l_i^t = l^t$, the given the allocations (x_i^1, l_i^1) of (x^1, l^1) under free trade one can find Pareto-optimal allocations (x_i^0, l_i^0) of (y^0, l^0) under autarky such that

$$p^1 \cdot x_i^1 - w^1 \cdot l_i^1 \geq p^1 \cdot x_i^0 - w^1 \cdot l_i^0 \text{ for } i = 1, 2, \ldots, N.$$

This proves that for *any* free-trade equilibrium it is possible to find a weakly Pareto-inferior Pareto-optimal autarky equilibrium. It does not prove the obverse proposition that for any autarky

equilibrium it is possible to find a weakly Pareto-superior free-trade equilibrium. A general gains-from-trade theorem was therefore yet to be established, but Samuelson had provided an important first step.

Hicks (1939) ushered in the 'new welfare economics' with a synthesis building on Hotelling (1938) and Kaldor (1939) and based on the compensation principle, making it possible, according to him, to make policy proposals in favour of economic efficiency which were free of value judgements. Hicks's most original contribution (Hicks 1940) was his attempt to apply the compensation principle to data on a country's real national income. This was a natural thing to try to do, since Pigou's (1920) main work was devoted to evaluating a country's welfare by national-income comparison, and it was largely Pigou's resort to interpersonal comparisons in order to justify this that was the object of Robbins's (1938) criticism.

Hicks's (1940) basic tool was the 'revealed-preference' comparison which had been employed by Barone (1908) and Hotelling (1938). If observations are available at times 0 and 1 of a country's national income in period-1 prices, and it is recorded that $p^1 \cdot y^1 \geq p^1 \cdot y^0$ (where p^t, y^t are vectors of prices and outputs at time t), what can be inferred? In the first place, to make any headway one must assume that the observed situations are competitive equilibria. Let us define an allocation of a commodity bundle x as an $N \times n$ matrix X whose ith row, x_i, is the bundle of n commodities allocated to individual i, and whose row sum $\sigma(X) = \sum_{i=1}^{N} x_i$ is equal to x. As between two bundles x_i^0, x_i^1 consumed by individual i, let us define $x_i^1 R_i x_i^0$ to mean that x_i^1 is preferred or indifferent to x_i^0 by individual i, where R_i is a continuous, convex, monotonic total order, with P_i denoting strict preference and I_i indifference. (This relation assumes the absence of externalities in consumption). Finally, let $X^1 R X^0$ (resp. $X^1 P X^0$) mean that X^1 is weakly (resp. strictly) Pareto-superior to X^0 (i.e. $x_i^1 R_i x_i^0$ for all i, resp. $x_i^1 R_i x_i^0$ for all i and $x_i^1 R_i x_i^0$ for some i). Then, from the real-income comparison $p^1 \cdot y^1 \geq p^1 \cdot y^0$, Hicks noted that there does not exist an allocation X of y^0 that is weakly Pareto-superior to the actual allocation X^1 of y^1. This follows from the same argument that establishes the Pareto optimality of the assumed competitive equilibrium in period 1. The non-existence of an allocation X of y^0 such that XRX^t, where $\sigma(X^1) = y^1$, constituted for Hicks the definition of an 'increase in real social income'.

Kuznets (1948) pointed out by an example that, in the case considered by Hicks, it could also be true that there is no allocation X of y^1 which is weakly Pareto superior to the actual allocation X^0 of y^0. Accordingly he suggested that Hicks's criterion be supplemented by the condition that there should exist an allocation $\overline{X}$ of y^1 that is weakly Pareto superior to the actual allocation X^0 of y^0. But while the latter criterion implies $p^0 \cdot y^1 \geq p^0 \cdot y^0$, it is not implied by it, so a national-income comparison using current and base prices would still not yield Kuznets's criterion.

Kuznets's criticism of Hicks was similar to the objection raised by Scitovsky (1941) to the criterion proposed by Kaldor (1939). According to Scitovsky's interpretation of Kaldor, an allocation X^1 of y^1 is better than an allocation X^0 of y^0, if there exists a reallocation $\overline{X}^1$ of y^1, which is Pareto superior to X^0. Scitovsky objected that this gave preference to the *status quo ante*, and besides, he pointed out that the criterion was internally inconsistent in the sense that it allowed two such pairs (X^t, y^t) to be superior to each other. He therefore proposed that Kaldor's test be supplemented by the criterion that there exist a reallocation $\overline{X}^0$ of y^0 that is Pareto inferior to X^1.

The literature on 'compensation tests' suffered from ambiguity as to the domain of definition of the relations and internal inconsistency of the relations. It was pointed out by Gorman (1955) that the relations were intransitive. It was shown in Chipman and Moore (1978) that the Hicks–Kuznets and Scitovsky double criteria, as well as the national-income comparisons in terms of base- and current-year prices, could lead to cycles of three competitive equilibria each superior to its successor.

The definitive contribution to the subject of national-income comparisons was that of Samuelson (1950) who introduced what Chipman and

Moore (1971) described as the 'Kaldor–Hicks–Samuelson (KHS) ordering'. The objects under comparison in this approach are sets Y of commodity bundles y, e.g. production-possibility sets. Letting $A(Y)$ denote the set of allocation matrices X such that $\sigma(X) \in Y$, this ordering is defined by

$$Y^1 >_R Y^R 0 \Leftrightarrow \left[\forall X^0 \in A\left(Y^0\right)\right]\left[\exists X^1 \in A\left(Y^1\right)\right] X^1 R X^0.$$

In words, Y^1, is potentially superior to Y^0 if, for all allocations of commodity bundles in Y^0, there exists a (weakly) Pareto superior allocation of a commodity bundle in Y^1. This is a reflexive and transitive relation; it also satisfies the condition that $Y^0 \subseteq Y^1$ implies $Y^1 >_R Y^0$. Samuelson also introduced the important concept of a utility- possibility frontier, which is the relative boundary of a utility-possibility set $U(Y, R; f)$; this in turn is a set of N-tuples of individual utilities, $u = f(X)$, for some $X \in A(Y)$, where f is an N-tuple of positive-valued utility functions representing R. If the sets Y are 'disposable' (that is, containing for every $y \in Y$ the bundles y' with $0 \leq y' \leq y$), and the R_i continuous and monotonic, then the utility-possibility sets are also disposable. If Y is non-empty, compact disposable, and convex, and the R_i are continuous, monotonic, and convex, then, provided the f_i are continuous and concave, $U(Y, R, f)$ is non-empty, compact, and convex (cf. Chipman and Moore 1971, p. 24). If the f_i are only quasi-concave and not concave, $U(Y, R, f)$ need not be convex (cf. Kannai and Mantel 1978). The KHS ordering among consumption-possibility sets translates into set-inclusion of the corresponding utility-possibility sets. Samuelson (1959, p. 10) gave an example of a case of crossing utility-possibility frontiers in which $X^2 \in A(Y^2)$ was Pareto superior to $X^1 \in A(Y^1)$ yet Y^1 would be ranked higher than Y^2 in terms of some value judgement. This established that the 'compensation tests' were not 'relatively *wertfrei*'.

Another approach was followed by Chipman and Moore (1973, 1976a), who asked the following question: if competitive equilibria (X^t, y^t, p^t) are observed satisfying $p^1 \cdot y^1 \geq p^1 \cdot y^0$ and $p^0 \cdot y^1 \geq p^0 \cdot y^0$, where $y^t \in Y_t$ for $t = 0, 1$, under what conditions on preferences must this imply that $Y^1 >_R Y^0$? For the case $Y^t = \{y^t\}$ they showed that the preference relations R_i must be identical and homothetic. This is a global result; with positive consumptions of all commodities the condition could no doubt be weakened to the aggregation criterion of Antonelli (1886), Gorman (1953), and Nataf (1953), namely, that consumer i's demand for commodity j have the form

$$x_{ij} = a_{ij}(p) + b_j(p) m_i$$

where m_i is consumer i's income.

Samuelson (1956) applied the compensation principle in a striking way in his proposed alternative to the new welfare economics. He discovered that, if a social- welfare function has the separable form W[$f(x)$], then a social utility function $f_w(x) = \max\{W[f(x)] : X \in A(x)\}$ has the property that it can be achieved in a decentralized manner by means of an income-distribution policy assigning individual shares of aggregate income as functions of prices and aggregate income. The first complete proof of this result was presented in Chipman and Moore (1972) (see also Chipman and Moore 1979; Chipman 1982). The main tool of analysis used was the concept of a Scitovsky indifference surface (Scitovsky 1942) which is defined as the boundary of the set $\sum_{i=1}^{N} R_i x_i$ where $R_i x_i$ is the set of all commodity bundles preferred or indifferent to x_i by individual i. This set is necessarily a subset of the set $R_w x$ of aggregate bundles preferred or indifferent to x by the Samuelson social ordering. In a competitive equilibrium the aggregate consumption bundle minimizes aggregate expenditure at the equilibrium prices over both sets, hence the bundle x_i minimizes each individual's expenditure over $R_i x_i$ (cf. Koopmans 1957, pp. 12–13).

Gains From Trade and Optimal Tariffs

The new tools developed by Scitovsky (1942) and Samuelson (1950, 1956) made possible a rigorous

proof of a gains-from-trade theorem, as well as of the proposition that a country could gain by a tariff.

Kemp (1962) noted that Samuelson's 1939 theorem implied that for any point on the free-trade utility-possibility frontier, the autarky utility-possibility frontier must pass below it; he reasoned that, as a result, for any point on the autarky utility-possibility frontier, the free-trade utility-possibility frontier must pass above it. If this argument can be accepted, it follows that for every allocation $X^0 \in A(Y^0)$ where Y^0 is the autarkic production-possibility set, there exists a (weakly) Pareto-superior allocation $X^1 \in A(Y^1)$ where Y^1 is the free-trade consumption-possibility set. Then free trade is superior to autarky by Samuelson's 1950 criterion.

The trouble with this argument, however, is that it requires that one can define a free-trade utility-possibility frontier (or consumption-possibility frontier) with the strong topological property of homeomorphism to the $(N-1)$-dimensional unit simplex (intuitively, absence of 'holes'). That this need not always be possible, was shown by Otani (1972, p. 149), and indeed admitted by Kemp and Wan (1972, p. 513). It is always possible if world prices are fixed, beyond our country's control. In that case the free-trade consumption-possibility set Y^1 is the budget set enclosing the production-possibility set Y^0 (cf. Samuelson 1962, p. 821), and the gains-from-trade theorem follows immediately from the property $Y^1 \supseteq Y^0 \Rightarrow Y^1 >_R Y^0$. In similar fashion the famous 'Baldwin envelope' (Baldwin 1948) defines a well-behaved consumption-possibility set containing the production-possibility set, from which one can prove the superiority of restricted trade (with an optimal tariff) to autarky (cf. Samuelson 1962).

For the general case in which a country can influence world prices, a method was shown by Kenen (1957). If all but 1 of the N individuals are constrained to have the same level of satisfaction under trade as achieved under autarky, a net production-possibility set can be constructed which indicates the amount available for the Nth person. It remains only to show that the Nth person will gain from a movement from autarky to free trade. A similar approach was indicated by Vanek (1964).

Grandmont and McFadden (1972) and Chipman and Moore (1972) both used the concept of an income-distribution policy to establish the gains-from-trade theorem. In Chipman and Moore this policy was chosen to be one that maximizes a separable Bergson-Samuelson social-welfare function. A standard argument is used to show that social utility is at least as high under free trade as under autarky. It remains to show that a function $W(u)$ can be chosen so that the corresponding distribution policy ensures that an increase in social utility implies an increase in each individual's utility. This is achieved by choice of $W(u) = min_i \left(u_i - u_i^0\right)/c_i$ where $c_i > 0$ and u_i^0 is the level of utility achieved by individual i under autarky.

General-Equilibrium Theory

The compensation principle is used in the proof of the theorem that every competitive equilibrium is Pareto-optimal (Arrow 1952, pp. 516, 519; Koopmans 1957, p. 49; Debreu 1959, pp. 94–5), in the sense that arbitrary allocations of feasible output bundles among consumers are assumed possible, regardless of resource-ownership constraints. A pair (X^0, p^0) is a competitive equilibrium for the production-possibility set Y if $X^0 R X$ for all $X \in A(Y)$ satisfying $Xp^0 \leq X^0 p^0$ and $y^0 p^0 \leq y p^0$ for all $y \in Y$, where $y^0 = \sigma(X^0) \in Y$. Pareto-optimality means that one cannot find an $X \in A(Y)$ such that XPX^0. The proof is by contradiction: XPX^0 implies $Xp^0 \geq X^0 p^0$ (the vector inequality being weak in all components and strict in at least one) hence taking column sums, $yp^0 > y^0 p^0$.

The converse theorem, that every Pareto optimum can be sustained by a competitive equilibrium, requires stronger assumptions which are awkward to state (cf. Arrow 1952, p. 518; Koopmans 1957, p. 50; Debreu 1959, p. 95). The basic idea of the proof (Koopmans 1957, pp. 50–52; Debreu 1959, p. 96) can be sketched in terms of the concept of a Scitovsky (1942) indifference surface. If X^0 is a Pareto-optimal

allocation for a closed, convex production-possibility set Y, then the interior of the Scitovsky set of X^0 can be written $p_k x_k^0 + \sum_{i \neq k} R_i x_i^0$ for some k.

Defining the allocation X^1 by $x_k^1 p_k x_k^0$ and $x_i^1 R_i x_i^0$ for $i \neq k$ we have $X^1 \; PX^0$ hence $X^1 \notin A(Y)$. Therefore the interior of the Scitovsky set does not intersect Y, and these convex sets can be separated by a hyperplane defining the equilibrium prices. It is then verified that at these prices the properties of a competitive equilibrium are satisfied.

Debreu (1954, p. 590) introduced an alternative equilibrium concept according to which the condition that consumer preferences be maximized subject to their budget constraints was replaced by the condition that consumer expenditures be minimized subject to the constraints that the bundles considered be at least as desirable as the equilibrium bundles. (The second of the above theorems follows more easily under this alternative definition). For a given set of positive-valued utility functions representing consumer preferences, Arrow and Hahn (1971, p. 108) called this a 'compensated equilibrium'. As a means of proving existence of the latter they studied the utility-possibility frontier or 'Pareto frontier' (1971, p. 96), and obtained a new proof of the result of Chipman and Moore (1971) that the set of Pareto-optimal allocations X of Y (the 'contract curve') and the utility-possibility frontier are topologically homeomorphic to the unit simplex of dimension one less than the number of individuals. These results were further developed by Moore (1975).

Cost–Benefit Analysis

Hicks (1941, p. 112) made an interesting distinction between two tasks of welfare economics: (1) the study of (Pareto)-optimal organizations of the economy and (2) the study of deviations from such optima. More precisely, the first was concerned with when there was a deviation and the second with the size of the deviation. He also identified these two tasks with general- and partial-equilibrium analysis respectively, although there appears to be no justification for this other than the historical accident that consumers' surplus developed as a partial-equilibrium tool. He remarked that consumers' surplus is not needed for the first task, since lack of fulfilment of the proportionality between marginal utilities and marginal costs provides the needed information immediately. For the second task, he was not content with a ranking of the non-optimal states, but with measuring the size of their deviations from optimality, which of course would provide such a ranking. Thus, the staunch ordinalist in consumer theory became an equally ardent cardinalist in consumer theory.

Hicks's concepts of compensating and equivalent variation (Hicks 1942) may most conveniently be defined in terms of the minimum-income or income-compensation functions of McKenzie (1957) and Hurwicz and Uzawa (1971). Denoting the ith consumer's demand function by $x_i = h_i(p, m_i)$ (where x_i and p are n-vectors), and defining the indirect preference relation R_i^* by $(p^0, m_i^0)\, R_i^* (p^1, m_i^1)$ if and only if $h_i(p^0, m_i^0) R_i h_i(p^1, m_i^1)$, the income-compensation function is defined by

$$\mu_i(p; p^0, m_i^0) = \inf\{m_i : (p, m_i) R_i^* (p^0, m_i^0)\}.$$

Following Chipman and Moore (1980b), the generalized compensating variation in going from (p^0, m_i^0) to (p, m_i) is defined as

$$C_i(p, m_i; p^0, m_i^0) = m_i - \mu_i(p; p^0, m_i^0)$$

and the generalized equivalent variation by

$$E_i(p, m_i; p^0, m_i^0) = \mu_i(p^0; p, m_i) - m_i^0.$$

These reduce to Hicks's concepts when $m_i = m_i^0$.

The compensating variation expresses for each consumer the amount of money income he or she would be willing to give up (or the negative of the amount by which he or she would have to be compensated), at the new prices, to make up for the change in prices and income. One of the reasons for the great appeal of the concept is that these are amounts that can be added up over the set of consumers. In Hicks's words (1942, p. 127):

C

> the general test for a particular reform being an *improvement* is that the gainers should gain sufficiently for them to be able to compensate the losers and still remain gainers on balance. This test would be carried out by striking the balance of the Compensating Variations.

Denoting by m' the vector of N incomes in state t, and by M^t their sum, we can define a dual potential-improvement ordering between pairs of price-income pairs (p^t, M^t) as follows. Let $A^*(p, M)$ be the set of $(n + N)$-tuples (p, m) such that $\sum_{i=1}^{N} m_i = M$, and let R^* be the relation such that (p^0, m^0) $R^*(p^1, m^1)$ if and only if $(p^0, m_i^0) R_i^* (p^1, m_i^1)$ for $i = 1, 2, \ldots, N$. Then we define the dual KHS relation $>_R{}^*$ by

$$(p^0, M^0) >_{R^*} (p^1, M^1) \Leftrightarrow \left[\forall (p', m') \in A^*(p^1, M^1)\left(\exists (p, m) \in A^*(p^0, M^0)\right)\right] : (p, m) R^* (p', m').$$

Choosing price-income pairs (p^0, m^0) and (p^1, m^1) satisfying this definition, since $\mu_i(p^t; p, m_i)$ is an indirect utility function representing R_i^* for $t = 0$ or 1, we have $\mu_i(p^t; p^0, m_i^0) \geq \mu_i(p^t; p^1, m_i^1)$ for all individuals i, hence

$$M^0 = \sum_{i=1}^{N} m_i^0 = \sum_{i=1}^{N} \mu_i(p^0; p^0, m_i^0) \geqq \sum_{i=1}^{N} \mu_i(p^0; p^1, m_i^1)$$

so one obtains a multi-consumer analogue to the compensating variation from the formula

$$M^0 = \sum_{i=1}^{N} \mu_i(p^0; p^1, m_i^1) \geqq$$

$$M^1 - \sum_{i=1}^{N} \mu_i(p^1; p^1, m_i^1) = 0.$$

Likewise for the equivalent variation,

$$0 = \sum_{i=1}^{N} \mu_i(p^0; p^0, m_i^0) - M^0 \geqq \sum_{i=1}^{N} \mu_i(p^0; p^1, m_i^1) - M^0.$$

In the latter case the same indirect utility functions are summed on both sides of the inequality sign; it is a case where Benthamites and compensationists can find common ground.

Boadway (1974) considered the relationship between the condition of positive summed compensating variations and the fulfilment of compensation tests and came to the negative conclusion that the former was neither necessary nor sufficient for satisfaction of the latter in general, but was sufficient in the case of identical and homothetic preferences. Foster (1976) showed that, if there are no price distortions (but not otherwise), satisfaction of the compensation tests implies satisfaction of the 'cost–benefit criterion' (positive summed compensating variations). This conclusion is in accord with the above inequalities.

What about the Hicksian tenet that the size of the compensating variation is important so that one can compare two suboptimal states? This would require one to be able to conclude that, if the compensating variation from state 0 to state 2 is positive and greater than the compensating variation from state 0 to state 1, then state 2 should be superior to state 1 in terms of the dual KHS ordering. But this is not true even in the case of the single consumer. It was shown in Chipman and Moore (1980) that the function $C_i(p, m_i; p^0, m_i^0)$ cannot be an indirect utility function for unrestricted domain $(p, m_i) > 0$, and can be if m_i is held constant if and only if preferences are homothetic, and if p_1 is held constant if and only if preferences are 'parallel' with respect to commodity 1. If preferences are identical and homothetic, since $\mu_i = \mu$ is homogeneous of degree 1 in m_i, $\sum_{i=1}^{N} \mu_i(p^0; p, m_i) = \mu(p^0, p, M)$, so exact aggregative analogues are obtained to both the

compensating and equivalent variations. If the equivalent variation, which is an indirect utility function, is used, restrictions on consumer preferences are not needed, and the problem of finding an adequate indicator of the size of the deviation from a given Pareto optimum is satisfactorily resolved.

Game Theory

One of the striking aspects of von Neumann and Morgenstern's theory of games (1947) was not only its postulate of measurability of utility but also that of its transferability between players. Since this was introduced as a positive rather than a normative assumption, it has met with even greater resistance of the part of economists than the hedonist calculus. Indeed, it was not until Debreu and Scarf (1963) showed how game theory could be liberated from this restriction with their development of the concept of the core of an economy that game theory began to be taken really seriously by economists. The replacement of transferability of utilities by transferability of commodities bears a striking resemblance to the replacement in welfare economics of the calculus of utilities by the principle of compensation.

In some branches of game theory the assumption of transferable utility is still retained, but it has been made somewhat more plausible, or at least interpretable, by means of the postulate that the utility functions of all individuals are linear in one distinguished commodity used for making side payments (cf. Owen 1982, p. 122). These utility functions have the form

$$U_i(x_{i1}, x_{i2}, \ldots, x_{in}) = c_i x_{i1} + V_i(x_{i2}, \ldots, x_{in}).$$

This form of the utility function goes back to Edgeworth (1891, p. 237n) and even earlier (though in garbled form) to Auspitz and Lieben (1889, p. 471). In Edgeworth it was used to illustrate the phenomenon of exchange when the marginal utility of one commodity serving as money was held constant, in accordance with one possible interpretation of Marshall's theory of consumers' surplus. (In the case $n = 2$ he showed that the exchange in commodity 2 would be constant, but in commodity 1 'indeterminate'; see the reply by Berry 1891, on behalf of Marshall, and Marshall 1891, p. 756; 1920, p. 845). The above form for the utility function has been rediscovered many time, by Wilson (1939), Samuelson (1942), and others; cf. Chipman and Moore (1976b, p. 115). Barone (1894, p. 213n) gave the name 'ideal money (numéraire)' to a good with a constant marginal utility (commodity 1 in the above). For the case $c_i = c$ for all i, these 'parallel' preferences (cf. Boulding 1945) yield a special case of the family of aggregable Antonelli–Gorman–Nataf demand functions referred to above.

Concluding Observations

As Scitovsky (1941) pointed out, the compensation principle has been used in two quite different ways. Prior to Hicks (1940), it was used only to compare efficient with inefficient states of a given economy with a given technology or trading system. Starting with Hicks (1940), its use was extended to comparison of efficient states of an economy under different technologies. It has turned out that, in order for national-income comparisons to provide a correct indicator of potential-welfare improvement, very strong conditions are required concerning similarity of individual preferences: locally, the Antonelli– Gorman–Nataf conditions, and, globally, identical homothetic preferences. It is not even enough to assume that aggregate demand can be generated by an aggregate preference relation – for example, that preferences are homothetic and relative income-distribution constant (cf. Chipman and Moore 1980a). Even in such cases, strong value judgements (such as acceptance of a particular Bergson–Samuelson social-welfare function) are required in order to draw welfare conclusions from national-income comparisons.

When attention is restricted to the efficient operation of an economy with a given technology, it turns out that, in most cases of interest, the ranking of consumption-possibility criterion sets according to the Kaldor–Hicks–Samuelson criterion follows from their ranking by set-inclusion.

This does not mean, however, that the set-inclusion is always obvious or easy to prove.

The KHS ordering of consumption-possibility sets could be given simply a factual interpretation as indicating the 'productive potential' of an economy. But if it is given a normative interpretation then it obviously involves a value judgement, since a more efficient outcome, if it is not Pareto-superior, can obviously be judged worse in terms of some social-welfare function.

Samuelson's (1956) model of the 'good society', elegant though it is, is too sweeping for most economists to accept, and it begs the question of how the social-welfare function will be chosen. Little's (1950) and Mishan's (1969) attempts to link plausible distributional value judgements with compensation criteria have encountered unresolvable logical difficulties (cf. Chipman and Moore 1978). The hope that the compensation principle would allow policy decisions to be made free of value judgements has not been fulfilled. Nevertheless, much has been learned about the interrelationships among values, facts, and policies, and it can certainly be said that the development of the compensation principle has led to clearer thinking about economic policy issues.

See Also

- ▶ Social Welfare Function
- ▶ Welfare Economics

Bibliography

Antonelli, G.B. 1886. *Sulla teoria matematica della economia politica.* Pisa: Folchetto. English translation: On the mathematical theory of political economy. In *Preferences, utility, and demand,* ed. J.S. Chipman, L. Hurwicz, M.K. Richter, and H.F. Sonnenschein. New York: Harcourt Brace Jovanovich, 1971.

Arrow, K.J. 1952. An extension of the basic theorems of classical welfare economics. In *Proceedings of the second Berkeley symposium on mathematical statistics and probability*, ed. J. Neyman. Berkeley/Los Angeles: University of California Press.

Arrow, K.J., and F.H. Hahn. 1971. *General competitive analysis*. San Francisco: Holden-Day.

Arrow, K.J., and T. Scitovsky, eds. 1969. *Readings in welfare economics*. Homewood: Irwin.

Auspitz, R., and R. Lieben. 1889. *Untersuchungen über die Theorie des Preises*. Leipzig: Duncker & Humblot.

Baldwin, R.E. 1948. Equilibrium in international trade: A diagrammatic analysis. *Quarterly Journal of Economics* 62: 748–762.

Baldwin, R.E. 1954. A comparison of welfare criteria. *Review of Economic Studies* 21: 154–161.

Barone, E. 1894. Sulla 'consumers' rent'. *Giornale degli Economisti.* Series 2, 9, September, 211–224.

Barone, E. 1908. Il Ministerio della produzione nello stato colletivista. *Giornale degli Economisti* Series 2. 37: 267–293; 391–414. English translation: The ministry of production in the collectivist state. In *Collectivist economic planning,* ed. F.A. Hayek. London: Routledge & Kegan Paul, 1935.

Berry, A. 1891. Alcune brevi parole sulla teoria del baratto di A. Marshall. *Giornale degli Economisti*. Series 2, 2, June, 549–553.

Boadway, R.W. 1974. The welfare foundations of cost-benefit analysis. *Economic Journal* 84: 926–939. A reply, *Economic Journal* 86 (1976), 358–361.

Boulding, K.E. 1945. The concept of economic surplus. *American Economic Review* 35: 851–869.

Cairnes, J.E. 1874. *Some leading principles of political economy newly expounded*. New York: Harper & Brothers.

Chipman, J.S. 1976. The paretian heritage. *Revue européenne des sciences sociales et Cahiers Vilfredo Pareto* 14 (37): 65–171.

Chipman, J.S. 1982. Samuelson and welfare economics. In *Samuelson and neoclassical economics*, ed. G.R. Feiwel. Boston: Kluwer-Nijhoff Publishing.

Chipman, J.S., and J.C. Moore. 1971. The compensation principle in welfare economics. In *Papers in quantitative economics*, ed. A.M. Zarley, vol. 2. Lawrence: University Press of Kansas.

Chipman, J.S., and J.C. Moore. 1972. Social utility and the gains from trade. *Journal of International Economics* 2: 157–172.

Chipman, J.S., and J.C. Moore. 1973. Aggregate demand, real national income, and the compensation principle. *International Economic Review* 14: 152–181.

Chipman, J.S., and J.C. Moore. 1976a. Why an increase in GNP need not imply an improvement in potential welfare. *Kyklos* 29 (3): 391–418.

Chipman, J.S., and J.C. Moore. 1976b. The scope of consumer's surplus arguments. In *Evolution, welfare, and time in economics*, ed. A. Tang, F.M. Westfield, and J.S. Worley. Lexington: Heath.

Chipman, J.S., and J.C. Moore. 1978. The new welfare economics, 1939–1974. *International Economic Review* 19: 547–584.

Chipman, J.S., and J.C. Moore. 1979. On social welfare functions and the aggregation of preferences. *Journal of Economic Theory* 21: 111–139.

Chipman, J.S., and J.C. Moore. 1980a. Real national income with homothetic preferences and a fixed distribution of income. *Econometrica* 48: 401–422.

Chipman, J.S., and J.C. Moore. 1980b. Compensating variation, consumer's surplus, and welfare. *American Economic Review* 70: 933–949.

Coase, R.H. 1960. The problem of social cost. *Journal of Law and Economics* 3: 1–44.

de Graaff, J. V. 1949. On optimum tariff structures. *Review of Economic Studies* 17: 47–59. Reprinted in Arrow and Scitovsky (1969).

de Graaff, J.V. 1957. *Theoretical welfare economics*. Cambridge: Cambridge University Press.

Debreu, G. 1951. The coefficient of resource utilization. *Econometrica* 19: 273–292.

Debreu, G. 1954. Valuation equilibrium and Pareto optimum. *Proceedings of the National Academy of Sciences* 40: 588–592.

Debreu, G. 1959. *Theory of value*. New York: Wiley.

Debreu, G., and H. Scarf. 1963. A limit theorem on the core of an economy. *International Economic Review* 4: 235–246.

Dupuit, J. 1844. De la mesure de l'utilité des travaux publics. *Annales des Ponts et Chaussées, Mémoires et documents relatifs à l 'art des constructions et au service de l'ingénieur.* Series 2, 2, 2e semestre, 332–375, Pl. 75. English translation: On the measurement of the utility of public works, in Arrow and Scitovsky (1969).

Edgeworth, F.Y. 1891. Osservazioni sulla teoria matematica dell'economia politica con riguardo speciale ai principi di economia di Alfredo Marshall. *Giornale degli Economisti.* Series 2, 2: 233–245. Ancora a proposito della teoria del baratto. *Giornale degli Economisti.* Series 2, 2: 316–318. Abridged English translation: On the determinateness of economic equilibrium, in F.Y. Edgeworth, *Papers relating to political economy*, vol. 2. London: Macmillan., 1925.

Foster, E. 1976. The welfare foundations of cost-benefit analysis – A comment. *Economic Journal* 86: 353–358.

Frisch, R. 1939. The Dupuit taxation theorem. *Econometrica* 7: 145–150. A further note on the Dupuit taxation theorem. *Econometrica* 7: 156–157.

Gorman, W.M. 1953. Community preference fields. *Econometrica* 21: 63–80.

Gorman, W.M. 1955. The intransitivity of certain criteria used in welfare economics. *Oxford Economic Papers,* N.S. 7: 25–35.

Grandmont, J.M., and D. McFadden. 1972. A technical note on classical gains from trade. *Journal of International Economics* 2: 109–125.

Harrod, R.F. 1938. Scope and method of economics. *Economic Journal* 48: 383–412.

Hicks, J.R. 1939. The foundations of welfare economics. *Economic Journal* 49: 696–712.

Hicks, J.R. 1940. The valuation of social income. *Economica,* N.S. 7: 105–124.

Hicks, J.R. 1941. The rehabilitation of consumers' surplus. *Review of Economic Studies* 8: 108–116. Reprinted in Arrow and Scitovsky (1969).

Hicks, J.R. 1942. Consumers' surplus and index-numbers. *Review of Economic Studies* 9: 126–137.

Hicks, J.R. 1957. *A revision of demand theory.* Oxford: Clarendon Press.

Hotelling, H. 1938. The general welfare in relation to problems of taxation and of railway and utility rates. *Econometrica* 6: 242–269. Reprinted in Arrow and Scitovsky (1969).

Hotelling, H. 1939. The relation of prices to marginal costs in an optimum system. *Econometrica* 7: 151–155. A final note, *Econometrica* 7: 158–159.

Hurwicz, L., and H. Uzawa. 1971. On the integrability of demand functions. In *Preferences, utility, and demand*, ed. J.S. Chipman, L. Hurwicz, M.K. Richter, and H.F. Sonnenschein. New York: Harcourt Brace Jovanovich.

Kaldor, N. 1939. Welfare propositions in economics and interpersonal comparisons of utility. *Economic Journal* 49: 549–552. Reprinted in Arrow and Scitovsky (1969).

Kaldor, N. 1940. A note on tariffs and the terms of trade. *Economica,* N.S. 7: 377–380.

Kannai, Y., and R. Mantel. 1978. Non-convexifiable Pareto sets. *Econometrica* 46: 571–575.

Kemp, M.C. 1962. The gains from international trade. *Economic Journal* 72: 803–819.

Kemp, M.C., and H.Y. Wan Jr. 1972. The gains from free trade. *International Economic Review* 13: 509–522.

Kenen, P.B. 1957. On the geometry of welfare economics. *Quarterly Journal of Economics* 71: 426–447.

Koopmans, T.C. 1957. *Three essays on the state of economic science*. New York: McGraw-Hill.

Kuznets, S. 1948. On the valuation of social income – Reflections on professor Hicks' article. *Economica,* N.S. 15: 1–16, 116–131.

Lerner, A.P. 1934. The concept of monopoly and the measurement of monopoly power. *Review of Economic Studies* 1: 157–175.

Little, I.M.D. 1950. *A critique of welfare economics*. 2nd ed. London: Oxford University Press. 1957.

Marshall, A. 1890. *Principles of economics.* London: Macmillan. 2nd ed, 1891; 8th ed, 1920.

McKenzie, L.W. 1957. Demand theory without a utility index. *Review of Economic Studies* 24: 185–189.

Mishan, E.J. 1969. *Welfare economics: An assessment*. Amsterdam: North-Holland.

Moore, J.C. 1975. The existence of 'compensated equilibrium' and the structure of the Pareto efficiency frontier. *International Economic Review* 16: 267–300.

Nataf, A. 1953. Sur des questions d'agrégation en économétrie. *Publications de l'Institut de Statistique de l'Université de Paris* 2 (4): 5–61.

Olsen, E. 1958. Udenrigshandelens gevinst [The gains of international trade]. *Nationalokonomisk Tiddskrift* 98 (1–2): 76–79.

Otani, Y. 1972. Gains from trade revisited. *Journal of International Economics* 2: 127–156.
Owen, G. 1982. *Game theory,* 2nd ed. Orlando: Academic Press.
Pareto, V. 1894. Il massimo di utilità dato dalla libera concorrenza. *Giornale degli Economisti* Series 2, 9: 48–66.
Pareto, V. 1895. Teoria matematica del commercio internazionale. *Giornale degli Economisti* Series 2, 10: 476–498.
Pareto, V. 1896–7. *Cours d'économie politique,* vols. 2. Lausanne: F. Rouge.
Pigou, A.C. 1920. *The economics of welfare,* 4th ed. London: Macmillan. 1932.
Ricardo, D. 1815. *An essay on the influence of a low price of corn on the profits of stock.* London: John Murray. In *The works and correspondence of David Ricardo,* ed. P. Sraffa, vol. 4. Cambridge: Cambridge University Press, 1951.
Robbins, L. 1938. Interpersonal comparisons of utility: A comment. *Economic Journal* 48: 635–641.
Samuelson, P.A. 1939. The gains from international trade. *Canadian Journal of Economics and Political Science* 5: 195–205.
Samuelson, P.A. 1942. Constancy of the marginal utility of income. In *Studies in mathematical economics and econometrics in memory of Henry Schultz*, ed. O. Lange, F. McIntyre, and T.O. Yntema. Chicago: University of Chicago Press.
Samuelson, P.A. 1947. *Foundations of economic analysis*. Cambridge, MA: Harvard University Press.
Samuelson, P.A. 1950. Evaluation of real national income. *Oxford Economic Papers,* N.S. 1: 1–29. Reprinted in Arrow and Scitovsky (1969).
Samuelson, P.A. 1956. Social indifference curves. *Quarterly Journal of Economics* 70: 1–22.
Samuelson, P.A. 1962. The gains from international trade once again. *Economic Journal* 72: 820–829.
Sanger, C.P. 1895. Recent contributions to mathematical economics. *Economic Journal* 5: 113–128.
Scitovsky, T. 1941. A note on welfare propositions in economics. *Review of Economic Studies* 9: 77–88. Reprinted in Arrow and Scitovsky (1969).
Scitovsky, T. 1942. A reconsideration of the theory of tariffs. *Review of Economic Studies* 9: 89–110.
Silberberg, E. 1980. Harold Hotelling and marginal cost pricing. *American Economic Review* 70: 1054–1057.
Vanek, J. 1964. A rehabilitation of 'well-behaved' social indifference curves. *Review of Economic Studies* 31: 87–89.
Viner, J. 1937. *Studies in the theory of international trade*. New York: Harper & Brothers.
von Neumann, J. and Morgenstern, O. 1947. *Theory of games and economic behavior*, 2nd ed. Princeton: Princeton University Press.
Wilson, E.B. 1939. Pareto versus Marshall. *Quarterly Journal of Economics* 53: 645–650.

Competing Risks Model

Gerard J. van den Berg

Abstract

A competing risks model is a model for multiple durations that start at the same point in time for a given subject, where the subject is observed until the first duration is completed and one also observes which of the durations is completed first. This article gives an overview of the main issues in the empirical econometric analysis of competing risks models. The central problem is the non-identification of dependent competing risks models. Models with regressors can overcome this problem, but it is advisable to include additional data. Alternatively, effects of interest can be bounded.

Keywords

Bounds; Censoring; Competing risks models; Copulas; Duration models; Hazard rates; Identification; Latent durations; Marriage and divorce; Maximum likelihood; Mixed proportional hazard models; Multiple-spell competing risks data; Multivariate duration models; Nonparametric kernel estimators; Regressors; Roy models; Selection; Semi-parametric models; Stochastic dominance; Unemployment durations; Unobserved heterogeneity; Weibull distributions

JEL Classifications

C13; C51

A competing risks model is a model for multiple durations that start at the same point in time for a given subject, where the subject is observed until the first duration is completed and one also observes which of the multiple durations is completed first.

The term 'competing risks' originates in the interpretation that a subject faces different risks

i of leaving the state it is in, each risk giving rise to its own exit destination, which can also be denoted by i. One may then define random variables T_i describing the duration until risk i is materialized. Only the smallest of all these durations $Y := \min_i T_i$ and the corresponding actual exit destination, which can be expressed as $Z := \text{argmin}_i T_i$, are observed. The other durations are censored in the sense that all is known is that their realizations exceed Y. Often those other durations are latent or counterfactual, for example if T_i denotes the time until death due to cause i.

In economics, the most common application concerns individual unemployment durations. One may envisage two durations for each individual: one until a transition into employment occurs, and one until a transition into non-participation occurs. We observe only one transition, namely, the one that occurs first. Other applications include the duration of treatments, where the exit destinations are relapse and recovery, and the duration of marriage, where one risk is divorce and the other is death of one of the spouses. More generally, the duration until an event of interest may be right-censored due to the occurrence of another event, or due to the data sampling design. The duration until the censoring is then one of the variables T_i.

Sometimes one is interested only in the distribution of Y. For example, an unemployment insurance (UI) agency may be concerned only about the expenses on UI and not in the exit destinations of recipients. In such cases one may employ standard statistical duration analysis for empirical inference with register data on the duration of UI receipt. However, in studies on individual behaviour one is typically interested in one or more of the marginal distributions of the T_i. If these variables are known to be independent, then again one may employ standard duration analysis for each of the T_i separately, treating the other variables T_j ($j \neq i$) as independent right-censoring variables. But often it is not clear whether the T_i are independent. Indeed, economic theory often predicts that they are dependent, in particular if they can be affected by the individual's behaviour and individuals are heterogeneous. It may even be sensible from the individual's point of view to use their privately observed exogenous exit rates into destinations j as inputs for the optimal strategy affecting the exit rate into destination i ($i \neq j$) (see, for example, van den Berg 1990). Erroneously assuming independence leads to incorrect inference, and in fact the issue of whether the durations T_i are related is often an important question in its own right.

Unfortunately, the joint distribution of all T_i is not identified from the joint distribution of Y, Z, a result that goes back to Cox (1959). In particular, given any specific joint distribution, there is a joint distribution with independent durations T_i that generates the same distribution of the observable variables Y, Z. In other words, without additional structure, each dependent competing risks model is observationally equivalent to an independent competing risks model. The marginal distributions in the latter can be very different from the true distributions.

Of course, some properties of the joint distribution are identified. To describe these it is useful to introduce the concept of the hazard rate of a continuous duration variable, say W. Formally, the hazard rate at time t is $\theta(t) := lim_{dt \downarrow 0} \Pr(W \in [t, \ t + dt))/dt$. Informally, this is the rate at which the duration W is completed at t given that it has not been completed before t. The hazard rate is the basic building block of duration analysis in social sciences because it can be directly related to individual behaviour at t. The data on Y, Z allow for identification of the hazard rates of T_i at t given that $T \geq t$. These are called the 'crude' hazard rates. If the T_i are independent, then these equal the 'net' hazard rates of the marginal distributions of the T_i.

We now turn to a number of approaches that overcome the general non-identification result for competing risks models. In econometrics, one is typically interested in covariate or regressor effects. The main approach has therefore been to specify semi-parametric models that include observed regressors X and unobserved heterogeneity terms V. With a single risk, the most popular

duration model is the mixed proportional hazard (MPH) model, which specifies that $\theta(t|\ X = x, V) = \psi(t)\ \exp\ (x'\beta)V$ for some function $\psi(.)$. V is unobserved, and the composition of the survivors changes selectively as time proceeds, so identification from the observable distributions of $T|X$ is non-trivial. However, it holds under the assumptions that $X \perp\!\!\!\perp V$ and $\text{var}(X) > 0$ and some regularity assumptions (see van den Berg 2001, for an overview of results). With competing risks, the analogue of the MPH model is the multivariate MPH (MMPH) model. With two risks,

$$\theta_1(t|\,x,\ V) = \psi_1(t)\ \exp(x'\beta_1)V_1 \text{ and}$$
$$\theta_2(t|\,x,\ V) = \psi_2(t)\ \exp(x'\beta l)V_2.$$

where $T_1, T_2|X, V$ are assumed independent, so that a dependence of the durations given X is modelled by way of their unobserved determinants V_1 and V_2 being dependent. Many empirical studies have estimated parametric versions of this model, using maximum likelihood estimation.

The semi-parametric model has been shown to be identified, under only slightly stronger conditions than those for the MPH model (Abbring and van den Berg 2003). Specifically, $\text{Var}(X) > 0$ is strengthened to the condition that the vector X includes two continuous variables with the properties that (a) their joint support contains a non-empty open set in $\mathbb{R}^2$, and (b) the vectors $\tilde{\beta}_1$, $\tilde{\beta}_2$ of the corresponding elements of β_1 and β_2 form a matrix $(\tilde{\beta}_1,\ \tilde{\beta}_2)$ of full rank. Somewhat loosely, X has two continuous variables that are not perfectly collinear and that act differently on θ_1 and θ_2. Note that, with such regressors, one can manipulate $\exp(x'\beta_1)$ while keeping $\exp(x'\beta_2)$ constant. The two terms $\exp(x'\beta_i)$ are identified from the observable crude hazards at $t = 0$ because at $t = 0$ no dynamic selection due to the unobserved heterogeneity has taken place yet. Now suppose one manipulates x in the way described above. If $T_1,\ T_2|X$ are independent, then the observable crude hazard rate of T_2 at $t > 0$, given that $T1 \geq t$, does not vary along. But, if $T1;\ T2|X$ are dependent, then this crude hazard rate does vary along, for the following reason. First, changes in $\exp(x'\beta_1)$ affect the distribution of unobserved heterogeneity V_1 among the survivors at t, due to the well-known fact that V_1 and X are dependent conditional on survival (i.e. conditional on $T_1 \geq t > 0$) even though they are independent unconditionally. Second, if V_1 and V_2 are dependent, this affects the distribution of V_2 among the survivors at t, which in turn affects the observable crude hazard of T_2 at t given that $T_1 \geq t$. In sum, the variation in this crude hazard with $\exp(x'\beta_1)$ for given $\exp(x'\beta_2)$ is informative on the dependence of the durations. An analogous argument holds for the crude hazard rate corresponding to cause $i = 1$.

Note that identification is not based on exclusion restrictions of the sort encountered in instrumental variable analysis, which require a regressor that affects one endogenous variable but not the other. Here, all explanatory variables are allowed to affect both duration variables – they are just not allowed to affect the duration distributions in the same way. Identification with regressors was first established by Heckman and Honoré (1989), who considered a somewhat larger class of models than the MMPH model and accordingly imposed stronger conditions on the support of X.

Although the MPH model is identified from single-risk duration data where we observe a single spell per subject, there is substantial evidence that estimates are sensitive to misspecification of functional forms of model elements (see van den Berg 2001, for an overview). This implies that estimates of MMPH models using competing-risks data should also be viewed with caution. It is advisable to include additional data. For example, longitudinal survey data on unemployment durations subject to right-censoring can be augmented with register data or retrospective data not subject to censoring (see for example van den Berg et al. 1994). More in general, one may resort to 'multiple-spell competing risks' data, meaning data with multiple observations of Y, Z for each subject. For a given subject, such observations can be viewed as multiple independent draws from the subject-specific distribution of Y, Z, on the assumption that the unobserved heterogeneity terms V_1, V_2 are identical across the spells of the subject. Here, a subject can denote a single physical unit, like an individual, for which we observe

two spells in exactly the same state, or it can denote a set of physical units for which we observe one spell each. Multiple-spell data allow for identification under less stringent conditions than single-spell data. Abbring and van den Berg (2003) showed that such data identify models that allow for full interactions between the elapsed durations t and x in $\theta_i(t|\, x, V)$, and, indeed, allow the corresponding effects to differ between the first and the second spell. The assumptions on the support of X are similar to above. Fermanian (2003) developed a nonparametric kernel estimator of the Heckman and Honoré (1989) model.

Another approach to deal with non-identification of dependent competing risks models is to determine bounds on the sets of marginal and joint distributions that are compatible with the observable data. Peterson (1976) derived sharp bounds in terms of observable quantities. They are often wide. In case of the marginal distributions of two sub-populations distinguished by a variable X, the bounds associated with the different X may overlap, whether or not X (monotonically) affects (one of) the marginal distributions. With overlap, the causal effects of X cannot even be signed.

Bond and Shaw (2006) combined bounds with regressors. In the case of a single binary regressor, the only substantive assumption made is that there exist increasing functions g and h such that $T_1, T_2|X = 0$ equals $g(T_1), h(T_2)|X = 1$ in distribution. In words, the dependence structure is invariant to the values of the regressors, so the latter affect only the marginal distributions. Specifically, the copula (and therefore Kendall's τ) of the joint distribution is invariant to the value of X. The assumption is satisfied by the aforementioned competing risks models with regressors. Clearly, by itself the assumption is insufficient for point identification. The bounds concern the regressor effects on the marginal distributions. If it is assumed that X affects the marginal distributions of T_i in terms of first-order stochastic dominance, the bounds are sufficient to sign the effect of X on at least one of the marginal distributions (so, in case of MMPH models, also on at least one of the individual marginal distributions conditional on V).

We end this article by noting some connections between competing risks models and other models. First, they are related to switching regression models, or Roy models. For example, if $T_i|X, V$ in the MMPH model have Weibull distributions, then we can write $\log T_i = x_i\alpha_i + \varepsilon_i (i = 1, 2)$ (for example, van den Berg et al. 1994), where we observe T_i iff $T_i < T_j (j \neq i)$. Second, competing risks models are building blocks of multivariate duration models, notably models where one of the durations is always observed (for example, T_1 captures the moment of a treatment and T_2 is the observed duration outcome of interest).

We have considered only continuous-time duration variables T_i that have different realizations with probability 1. Recently, semi-parametric and nonparametric results have been derived for discrete-time or interval-censored competing risks models and models where different risks can be realized simultaneously (see for example Bedford and Meilijson 1997; van den Berg, van Lomwel and van Ours 2003; Honoré and Lleras-Muney 2006). The biostatistical literature contains many studies in which specific assumptions are made on the dependence structure of the two durations T_i, enabling inference on the marginal distributions from data on Y, Z (see for example Moeschberger and Klein 1995, for a survey).

See Also

- Partial Identification in Econometrics
- Proportional Hazard Model
- Selection Bias and Self-Selection

Bibliography

Abbring, J., and G. van den Berg. 2003. The identifiability of the mixed proportional hazards competing risks model. *Journal of the Royal Statistical Society, Series B* 65: 701–710.

Bedford, T., and I. Meilijson. 1997. A characterization of marginal distributions of (possibly dependent) lifetime variables which right censor each other. *Annals of Statistics* 25: 1622–1645.

Bond, S., and J. Shaw. 2006. Bounds on the covariate-time transformation for competing-risks survival analysis. *Life time Data Analysis* 12: 285–303.

Cox, D. 1959. The analysis of exponentially distributed life-times with two types of failure. *Journal of the Royal Statistical Society, Series B* 21: 411–421.
Fermanian, J. 2003. Nonparametric estimation of competing risks models with covariates. *Journal of Multivariate Analysis* 85: 156–191.
Heckman, J., and B. Honoré. 1989. The identifiability of the competing risks model. *Biometrika* 76: 325–330.
Honoré, B., and A. Lleras-Muney. 2006. Bounds in competing risks models and the war on cancer. *Econometrica* 74: 1675–1698.
Moeschberger, M., and J. Klein. 1995. Statistical methods for dependent competing risks. *Lifetime Data Analysis* 1: 195–204.
Peterson, A. 1976. Bounds for a joint distribution function with fixed sub-distribution functions: application to competing risks. *Proceedings of the National Academy of Sciences* 73: 11–13.
van den Berg, G. 1990. Search behaviour, transitions to nonparticipation and the duration of unemployment. *Economic Journal* 100: 842–865.
van den Berg, G. 2001. Duration models: Specification, identification, and multiple durations. In *Handbook of econometrics*, ed. J. Heckman and E. Leamer, Vol. 5. Amsterdam: North-Holland.
van den Berg, G., M. Lindeboom, and G. Ridder. 1994. Attrition in longitudinal panel data, and the empirical analysis of dynamic labour market behaviour. *Journal of Applied Econometrics* 9: 421–435.
van den Berg, G., van Lomwel, A., and van Ours, J. 2003. Nonparametric estimation of a dependent competing risks model for unemployment durations. Discussion Paper No. 898. Bonn: IZA.

Competition

George J. Stigler

Abstract

Competition arises whenever two or more parties strive for something that all cannot obtain. The classical economists felt no need for a very precise definition of competition because they viewed monopoly as highly exceptional. In the late 19th century competition became the subject of intense analysis; the concept of perfect competition emerged as the standard model of economic theory and as first approximation in the concrete studies of applied microeconomics. The limitations of the concept in dealing with conditions of persistent and imperfectly predicted change will be removed only when economics possesses a developed theory of change.

Keywords

Bilateral monopoly; Clark, J.; Cliffe Leslie, T.; Coalitions; Comparative statics; Competition; Competitive equilibrium; Contracting; Cournot, A.; Creative destruction; Darwin, C.; Demsetz, H.; Edgeworth, F.; Entrepreneurship; Factor price equalization theorem; Fisher, I.; Industrial organization; Innovation; Jenkin, F.; Jevons, W.; Joint action; Kirzner, I.; Knight, F.; Labour markets; Laissez faire; Law of indifference; Law of one price; Law of supply and demand; Long-run equilibrium; Malthus, T.; Market price; Markets; Marshall, A.; Mathematics and economics; Mill, J. S.; Monopoly; Oligopoly; Pareto, V.; Perfect competition; Perfect information; Perfect markets; Pigou, A.; Profit-maximixing behaviour; Rate of return; Recontracting; Resource mobility; Schumpeter, J.; Second best; Senior, N.; Smith, A.; Stigler, G. J.; Thornton, H.; Walras, L.; Workable competition

JEL Classifications

B0

Competition is a rivalry between individuals (or groups or nations), and it arises whenever two or more parties strive for something that all cannot obtain. Competition is therefore at least as old as man's history, and Darwin (who borrowed the concept from economist Malthus) applied it to species as economists had applied it to human behaviour.

A concept that is applicable to two cobblers or a thousand shipowners or to tribes and nations is necessarily loosely drawn. When Adam Smith launched economics as a comprehensive science in 1776, he followed this usage. He explained why a reduced supply of a good led to a higher price: the 'competition [which] will immediately begin' among buyers would bid up the price. Similarly if the supply become larger, the price would sink

more, the greater 'the competition of the sellers' (Smith [1776] 1976, pp. 73–4). Here competition was very much like a race: a race to obtain part of reduced supplies or to dispose of a part of increased supplies. Almost nothing except a number of buyers and sellers was necessary for competition to operate. And the greater the number of each, the greater the vigour of competition:

> If this capital [sufficient to trade in a town] is divided between two different grocers, their competition will tend to make both of them sell cheaper, than if it were in the hands of one only; and if it were divided among twenty, their competition would be just so much the greater, and the chance of their combining together, in order to raise the price, just so much the less. (ibid., pp. 361–2)

With such a loose concept, there was little occasion to speak of one market as being more or less competitive than another, although this very passage presented the commonsense idea that larger numbers of rivals increased the intensity of competition.

The competition of grocers in a town pertained to competition *within* a market or an industry. Smith made much of the competition of different markets or industries for resources, and he developed what has always remained the main theorem on the allocation of resources in an economy composed of private, competing individuals or enterprises. The argument may be stated: Each owner of a productive resource will seek to employ it where it will yield the largest return. As a result, under competition each resource will be so distributed that it yields the same rate of return in every use. For if a resource were earning more in one use than another, it would be possible for its return in the lower-yielding use to be increased by reallocating it to the higher-yielding use. And this theorem led to what John Stuart Mill called the most frequently encountered proposition in economics: 'There cannot be two prices in the same market' (Mill 1848, Book II, ch. IV, s. 3).

The competition of different markets or industries for the use of the same resources called attention to some problems which are less important within a single market such as the grocery trade in a town. One must possess knowledge of the investment opportunities in these different employments, and that knowledge is less commonly possessed than knowledge within one market. It often requires a good deal of time to disengage resources from one field and install them elsewhere. Both of these conditions were recognized by Smith, who spoke of the difficulty of keeping secret the existence of extraordinary profits, and of the long run sometimes required for the attainment of equality of rates of return.

For the next three-quarters of a century the prevailing treatment of competition followed the practice of Smith. One can find occasional hints of a more precise definition of competition, well illustrated by Nassau W. Senior:

> But though, under free competition, cost of production is the regulator of price, its influence is subject to much occasional interruption. Its operation can be supposed to be perfect only if we suppose that there are no disturbing causes, that capital and labour can be at once transferred, and without loss, from one employment to another, and that every producer has full information of the profit to be derived from every mode of production. But it is obvious that these suppositions have no resemblance to the truth. A large portion of the capital essential to production consists of buildings, machinery, and other implements, the results of much time and labour, and of little service for any except their existing purposes . . . few capitalists can estimate, except upon an average of some years, the amount of their own profits, and still fewer can estimate those of their neighbours. (1836, p. 102)

Senior is hinting at a concept of perfect competition, but the hint is not pursued.

The classical economists felt no need for a precise definition because they viewed monopoly as highly exceptional: Harold Demsetz has counted only one page in 90 devoted to monopoly in *The Wealth of Nations* and only one in 500 in Mill's *Principles of Political Economy*. Indeed the word 'monopoly' was usually restricted to grants by the sovereign of exclusive rights to manufacture, import or sell a commodity; witness the entry in the *Penny Cyclopedia* (1839):

> It seems then that the word monopoly was never used in English Law, except when there was a royal grant authorizing some one or more persons only to deal in or sell a certain commodity or article.
>
> If a number of individuals were to unite for the purpose of producing any particular article or commodity, and if they should succeed in selling such

> article very extensively, and almost solely, such individuals in popular language would be said to have a monopoly. Now, as these individuals have no advantages given them by the law over other persons, it is clear they can only sell more of their commodity than other persons by producing the commodity cheaper and better. (XV, p. 341)

The ability of rivals to seek out and compete away supernormal profits, unless prevented by legal obstacles, was believed to be the basic reason for the pervasiveness of competition.

In the last third of the 19th century the concept of competition became the subject of intense study. The most popular reason given for this attention is that the growth of large-scale enterprises, including railroads, public utilities, and finally great manufacturing enterprises, made obvious the fact that a simple concept of competition no longer fit the economy of an industrial nation such as England.

A second source of misgiving with the broad definition of competition is that it might not lead to the uniformity of returns to a resource predicted by the theory. The Irish economist Cliffe Leslie repeatedly made this charge:

> Economists have been accustomed to assume that wages on the one hand and profits on the other are, allowing for differences in skill and so forth, equalized by competition, and that neither wages nor profits can anywhere rise above 'the average rate', without a consequent influx of labour or of capital bringing things to a level. Had economists, however, in place of reasoning from an assumption, examined the facts connected with the rate of wages, they would have found, from authentic statistics, the actual differences so great, even in the same occupation, that they are double in one place what they are in another. Statistics of profits are not, indeed, obtainable like statistics of wages; and the fact that they are not so, that the actual profits are kept a profound secret in some of the most prominent trades, is itself enough to deprive the theory of equal profits of its base. (1888, pp. 158–9)

The easiest way to combat such criticisms was not to confront them with data – that path was not chosen for many years – but to define competition in such a way as to ensure the desired results such as uniformity of price.

The complications possible with competition were raised also on the theoretical side. William T. Thornton, in his book *On Labour* (1869), denied the fact that prices were determined by the 'law of supply and demand', particularly within labour markets. He employed bizarre examples, such as supply and demand curves which coincided over a vertical range, to show that price could be indeterminate or unresponsive to changes in supply or demand. These objections naturally called forth responses, from both J.S. Mill (*Collected Works*, V) and Fleeming Jenkin, a famous engineer.

The most persuasive reason for the increasing attention to the concepts of economics was the gradual move of economic studies to the universities, which proceeded rapidly in the last decades of the century. The expanding use of mathematics was one major symptom of the development of the formal and abstract theory of economics by Walras, Pareto, Irving Fisher and others. That formalization would scarcely be possible without a more precise specification of the nature of competition, and the precise specification of the nature of competition, and the replies to Thornton's criticisms were a precursor to this literature.

The groundwork for the development of the concept of perfect competition was laid by Augustin Cournot in 1838 in his *Mathematical Principles of the Theory of Wealth*. He made the first systematic use of the differential calculus to study the implications of profit-maximizing behaviour. Starting with the definition, Profits = Revenue − Costs, Cournot sought to maximize profits under various market conditions. He faced the question: How does revenue (say, pq) vary with output (q)? The natural answer is to *define* competition as that situation in which p does not vary with q – in which the demand curve facing the firm is horizontal. This is precisely what Cournot did:

> The effects of competition have reached their limit, when each of the partial productions Dk [the output of producer k] is *inappreciable*, not only with reference to the total production $D = F(p)$, but also with reference to the derivative $F'(p)$, so that the partial production Dk could be subtracted from D without any appreciable variation resulting in the price of the commodity. (Cournot [1838] 1927, p. 90)

This definition of competition was especially appropriate in Cournot's system because,

according to his theory of oligopoly, the excess of price over marginal cost approached zero as the number of like producers became large. The argument is as follows:

Let the revenue of the firm be q_ip, and let n identical firms have the same marginal costs, MC. Then the equation for maximum profits for one firm would be

$$p + q_i(\mathrm{d}p/\mathrm{d}q) = MC.$$

The sum of n such equations would be

$$np + q(\mathrm{d}p/\mathrm{d}q) = nMC,$$

for $nq_i = q$. This last equation may be written,

$$p = \mathrm{MC} - p/nE,$$

where E is the elasticity of market demand (Cournot 1838, p. 84).

Cournot believed that this condition of competition was fulfilled 'for a multitude of products, and, among them, for the most important products'.

Cournot's definition was enormously more precise and elegant than Smith's so far as the treatment of numbers was concerned. A market departed from unlimited competition to the extent that prices exceeded the marginal cost of the firm, and the difference approached zero as the number of rivals approached infinity. This definition, however, illuminated only the effect of number of rivals on the power of individual firms to influence the market price, on Cournot's special assumption that each rival believed that his output decisions did not affect the output decisions of his rivals. It therefore bore only on what we term market competition.

Cournot did not face the question of the role of information possessed by traders, and this question was taken up by William Stanley Jevons in 1871 in his *Theory of Political Economy*. He characterized a perfect *market* by two conditions:

> (1.) A market, then, is theoretically perfect only when all traders have perfect knowledge of the conditions of supply and demand, and the consequent ratio of exchange; . . . (2.) . . . there must be perfectly free competition, so that any one will exchange with any one else upon the slightest advantage appearing. There must be no conspiracies for absorbing and holding supplies to produce unnatural ratios of exchange. (Jevons 1871, pp. 86, 87)

By perfect knowledge Jevons meant only that each trader in a market knew the price bids of every other trader. The second condition ruled out any joint actions by two or more traders, without his noticing that with knowledge so perfect as to know the behaviour of rivals, there might appear the very conspiracies he ruled out. The two conditions dictated that 'there cannot be two prices for the same kind of article' in a perfect market, which he called the 'law of indifference'.

The merging of the concepts of competition and the market was unfortunate, for each deserved a full and separate treatment. A market is an institution for the consummation of transactions. It performs this function efficiently when every buyer who will pay more than the minimum realized price for any class of commodities succeeds in buying the commodity, and every seller who will sell for less than the maximum realized price succeeds in selling the commodity. A market performs these tasks more efficiently if the commodities are well specified and if buyers and sellers are fully informed of their properties and prices. Also a complete, perfect market allows buyers and seller to act on differing expectations of future prices. A market may be perfect and monopolistic or imperfect and competitive. Jevons's mixture of the two has been widely imitated by successors, of course, so that even today a market is commonly treated as a concept subsidiary to competition.

Edgeworth was the first economist to attempt a systematic and rigorous definition of perfect competition. His exposition deserves the closest scrutiny in spite of the fact that few economists of his time or ours have attempted to disentangle and uncover the theorems and conjectures of the *Mathematical Psychics* (1881), probably the most elusively written book of importance in the history of economics. His exposition was the most influential in the entire literature.

The conditions of perfect competition are stated as follows:

> The *field of competition* with reference to a contract, or contracts, under consideration consists of all individuals who are willing and able to recontract about the articles under consideration . . .
>
> There is free communication throughout a *normal* competitive field. You might suppose the constituent individuals collected at a point, or connected by telephones – an ideal supposition [1881], but sufficiently approximate to existence or tendency for the purposes of abstract science.
>
> A *perfect* field of competition professes in addition certain properties peculiarly favourable to mathematical calculation; . . . The conditions of a *perfect* field are four; the first pair referable to the heading *multiplicity* or continuity, the second to *dividedness* or fluidity.
>
> I. An individual is free to *recontract* with any out of an indefinite number, . . .
> II. Any individual is free to *contract* (at the same time) with an indefinite number;
>
> . . . This condition combined with the first appears to involve the indefinite divisibility of each *article* of contract (if any *X* deal with an indefinite number of *Y*s he must give each an indefinitely small portion of *x*); which might be erected into a separate condition.
>
> III. Any individual is free to *recontract* with another independently of, *without the consent* being required of, any third party, . . .
> IV. Any individual is free to *contract* with another independently of a third party; . . .

The failure of the first [condition] involves the failure of the second, but not vice versa; and the third and fourth are similarly related (Edgeworth 1881, pp. 17–19).

The essential elements of this formidable list of conditions are two:

1. There are an indefinitely large number of independent traders on each side of a market (the Cournot condition).
2. Each trader can costlessly make tentative contracts with everyone (hence the divisibility of commodities) and alter these contracts (recontract) so long as a more favourable contract can be made. The result is perfect knowledge (the Jevonian condition).

Edgeworth gave an intuitive argument for the need for an indefinitely large number of traders on both sides of a market. It proceeds as follows. Let there be one seller and two buyers, and let the seller gain all the benefits of the sale: each buyer is charged the maximum price he would pay rather than withdraw from the market. If now a second seller appears, he will find it advantageous to offer better terms to the two buyers: 'It will in general be possible for *one* of the [sellers] (without the consent of the other), to *recontract* with the two [buyers], so that for all those three parties the recontract is more advantageous than the previously existing contract' (ibid., p. 35). As the numbers of traders on each side increase, the price approaches the competitive equilibrium level where no individual trader can influence it.

A defect in this argument is that it ignores the fact that if the traders on one or both sides of the market, be they 2, or 2000 or 2,000,000, join together they can do better *individually* than by competing. If traders on each side join, however, there will be bilateral monopoly, not competition. Edgeworth gives no reason why the combination of traders fails to take place. Only in modern times has the reason for independent behaviour by rivals been established: the costs of reaching and enforcing agreements on joint action increase with both the number of rivals and the complexity of the transactions. At a certain level – quite possibly with only two traders under some conditions – the costs of joint action exceed the gain to at least some of the traders, and independent behaviour emerges.

Edgeworth's 'conjecture', as it is now often called, that a unique, competitive price would emerge when the number of traders became large, has given rise to a modern literature vast in scope and often highly advanced in its mathematical techniques (for references, see Hildenbrand 1974). One result in this literature is that in the case of a large (infinite) number of traders, no coalition of a portion of the traders can exclude traders outside the coalition from trading at the price-taking equilibrium.

Edgeworth's introduction of the requirement that the commodity or service that is traded be highly divisible is a response to the following problem:

> Suppose a market, consisting of an equal number of masters and servants, offering respectively wages and service; subject to the condition that no man can

> serve two masters, no master employ more than one man; or suppose equilibrium already established between such parties to be disturbed by any sudden influx of wealth into the hands of the masters. Then there is no *determinate*, and very generally *unique*, arrangement towards which the system tends under the operation of, may we say, a law of Nature, and which would be predictable if we knew beforehand the real requirements of each, or of the average, dealer; . . . (Edgeworth 1881, p. 46).

Consider the simple example: a thousand masters will each employ a man at any wage below 100; a thousand labourers will each work for any wage above 50. There will be a single wage rate: knowledge and numbers are sufficient to lead a worker to seek a master paying more than the going rate or a master to seek out a worker receiving less than the market rate. But any rate between 50 and 100 is a possible equilibrium. But if a single worker leaves the market, the wage will rise to 100, and if a single employer withdraws, the wage will fall to 50. This ability of a single trader to affect the price arises because of the lumpiness of the article traded (here a worker's labour for a given period). Once a worker can work for two masters, the withdrawal of one worker in a thousand will reduce the available hours of work per day to each employer by only 8/1000 hours or 4.8 minutes per day, with only negligible influence upon the wage rate. Alternatively, a distribution of wage offers and demands would also eliminate the indeterminacy and market power.

Edgeworth's analysis was limited to competition within a market, and it was left to John Bates Clark to emphasize the need for mobility of resources if the return on each resource was to be equalized in every use.

> . . .there is an ideal arrangement of the elements of society, to which the force of competition, acting on individual men, would make the society conform. The producing organism actually shapes itself about his model, and at no time does it vary greatly from it . . . We must use assumptions boldly and advisedly, make labour and capital absolutely mobile, and letting competition work in ideal perfection. (Clark 1899, pp. 68, 71)

Perfect and free mobility of resources is of course an even more extreme assumption than the other conditions required for perfect competition because there is less reason to believe that free movement of resources is even approached in the real economy. Nor is the assumption of perfect mobility necessary to eliminate monopoly power in a market: in the Victorian age, the price of wheat of Iowa was set in Liverpool even though transportation costs were substantial. The assumption is usually necessary to attain strict equality in the price of a good at every point (the law of one price), although even this is not strictly true (as in the factor price equalization theorem). Clark also demanded that the economy be stationary for perfect competition, a condition we shall return to later.

All the elements of a concept of perfect competition were in place by 1900, and this concept increasingly became the standard model of economic theory thereafter. The most influential statement of the conditions for perfect competition was made by Frank H. Knight in his doctoral dissertation, *Risk, Uncertainty and Profit* (1921). The conditions were stated in extreme form; for example, 'There must be perfect, continuous, costless intercommunication between all individual members of the society' (Knight 1921, p. 78) – so Jones in Seattle would know the price of potatoes and be able costlessly to ship to Smith in Miami a bushel of potatoes at every moment of time.

Of course these conditions are not *necessary*, but only sufficient, to achieve the competitive equilibrium. For example, if even a considerable fraction of buyers knows that seller A is charging more than B for a given commodity, their patronage may be quite enough to force A to reduce his price to that of B. Nor are the various conditions independent of one another: for example, if it is very cheap for either a commodity or its buyers or sellers to move between two places, that will insure that the prices in the two places will be widely known.

Along with the development of the concept of competition as a standard component of the theory of prices and the allocation of resources, it acquired a growing role as the criterion by which to judge the efficiency of actual markets. Adam Smith had already advanced the proposition that output was maximized in a private enterprise economy with competition. If each owner of a

resource maximized the return from his resources, then (in the absence of 'external' effects of one person's actions on others) aggregate output would be maximized. This theorem (labelled 'on maximum satisfaction') was developed and qualified by Léon Walras (1874), Alfred Marshall (1980), Pareto (1895–6, 1907), Pigou (1912) and a host of modern economists.

Competition is much too central a concept in economics to remain unaffected when economists change their interests or analytical methods. We may illustrate this fact by the problem of economic change.

In a regime of change, of growing population and capital or innovations or new consumer demands, the problem of defining competition is much more difficult than it is for the stationary economy. Unless the change is predictable with precision, knowledge must necessarily be incomplete and errors and lags in adaptation to new conditions can be large. For this reason, indeed, J.B. Clark believed that perfect competition was achievable only in the stationary economy.

Even short-run changes in market price raise the question: is the change in price initiated by a particular seller or buyer, and if so, is this trader not facing a negatively sloping demand curve or a positively sloping supply curve? The infinitely elastic supply and demand curves of perfectly competitive equilibrium seem inapplicable to periods of changing market conditions. Some economists nevertheless retain the condition that individual traders cannot influence price by introducing a hypothetical auctioneer who announces price changes.

A partial adaptation of the competitive concept to change is made by making it a long-run equilibrium concept. Even if resources are not costlessly mobile and even if entrepreneurs do not have perfect foresight, one can analyse the rate of approach of returns on resources to equality. If an industry experiences a once-for-all large change, it could be in competitive equilibrium before and after the change, and the equilibria could be studied by competitive theory (comparative statics).

This adaptation did not satisfy Joseph Schumpeter, who believed that incessant change in products and production methods was the very essence of competitive capitalism. He argued that the displacing of one product or method by another, a process which he called creative destruction, made the concept of perfect competition irrelevant to either positive analysis or welfare judgements. If the monopoly that reduced output, compared to competition, by 10 per cent in one year, increased output by 100 per cent over the next two decades, then monopoly might be preferred to stagnant competition.

It is crucial to this argument that monopoly provides large, though temporary, rewards to successful innovators but competition does not:

> But perfectly free entry into a *new* field may make it impossible to enter it at all. The introduction of new methods of production and new commodities is hardly conceivable with perfect – and perfectly prompt – competition from the start. And this means that the bulk of what we call economic progress is incompatible with it. As a matter of fact, perfect competition is and always has been temporarily suspended whenever anything new is being introduced – automatically or by measures devised for that purpose – even in otherwise perfectly competitive conditions. (Schumpeter 1942, pp. 104–5)

Schumpeter relies on instantaneous rivalry to eliminate the incentives to innovation under competition, and the conclusion would not hold if competition is defined in terms of long-run equilibrium.

Nevertheless the issue is not disposed of so easily. If change is continuous rather than sporadic, long-run equilibria will never be fully achieved. Several economists have emphasized that alterations in the concept of competition are called for in periods of historical change. Kirzner has emphasized the role of entrepreneurial rivalry in competition, whereas such rivalry is nonexistent in a perfectly competitive equilibrium. Demsetz has proposed a concept of laissez-faire competition, in which freedom of resources to move into any use is the central element. Such realistic reversions to the competitive concept of the classical economists have not been systematically formalized into theoretical models.

The concept of perfect competition, or indeed any theoretically precise concept of competition,

will not be met by the actual condition of competition in any industry. John Maurice Clark made the most influential effort to create a concept of 'workable competition' which would serve as a working rule for public policies which seek to preserve or increase competition.

Clark emphasized the fact that if one requisite of perfect competition is absent, it may be desirable that a second requisite also be unfulfilled. For example, with instantaneous mobility but imperfect knowledge, members of an occupation would keep shifting back and forth between two cities, always overshooting the amount of migration which would equalize wage rates. This propensity to overshoot equilibrium would be corrected with less mobility of labour. This problem was later formalized as the theory of the 'second best'.

The essence of the concept of workable competition was the belief that 'long- run curves, both of cost and of demand, are much flatter than short-run curves, and much flatter than the curves which are commonly used in the diagrams of theorists' (J.M. Clark 1940, p. 460). This correct and sensible view led to a proliferation of studies, usually in doctoral dissertations, of individual industries, in which the workableness of competition in each industry was appraised. Unfortunately there were no objective criteria to guide these judgements, and there was no evidence that the studies were accepted by the governmental agencies which administered competitive policies.

The popularity of the concept of perfect competition in theoretical economics is as great today as it has ever been. The concept is equally popular as first approximation in the more concrete studies of markets and industries that comprise the field of 'industrial organization' (applied microeconomics). The limitations of the concept in dealing with conditions of persistent and imperfectly predicted change will not be removed until economics possesses a developed theory of change. Even within a stationary economic setting the concept is being deepened by mathematical economists (see Mas-Colell 1982). Meanwhile the central elements of competition – the freedom of traders to use their resources where they will, and exchange them at any price they wish – will continue to play a major role in the economics of an enterprise economy.

See Also

- ▶ Exchange
- ▶ Large Economies
- ▶ Perfect Competition

Bibliography

Clark, J.B. 1899. *The distribution of wealth*. London: Macmillan.

Clark, J.M. 1940. Toward a concept of workable competition. *American Economic Review*, June; reprinted in *Readings in the social control of industry*. Philadelphia: Blakiston, 1942.

Cliffe Leslie, T.E. 1888. *Essays in political economy and moral philosophy*. London: Longmans Green.

Cournot, A. 1838. *Researches into the mathematical principles of the theory of wealth*. Reprinted. New York: Macmillan, 1927.

Demsetz, H. 1982. *Economic, legal and political dimensions of competition*. Amsterdam: North-Holland.

Edgeworth, F.Y. 1881. *Mathematical psychics*. London: Kegan Paul, 1932.

Hildenbrand, W. 1974. *Core and equilibria of a large economy*. Princeton: Princeton University Press.

Jevons, W.S. 1871. *The theory of political economy*. London: Macmillan.

Kirzner, I.M. 1973. *Competition and entrepreneurship*. Chicago: University of Chicago Press.

Knight, F.H. 1921. *Risk, uncertainty and profit*, Part 2. Boston: Houghton Mifflin Co.

Marshall, A. 1890. *Principles of economics*. London: Macmillan.

Mas-Colell, A., ed. 1982. *Noncooperative approaches to the theory of perfect competition*. New York: Academic Press.

McNulty, P.J. 1967. A note on the history of perfect competition. *Journal of Political Economy* 75: 395–399.

Mill, J.S. 1848. *Principles of political economy*. In *Collected works*, ed. J.M. Robson. Toronto: University of Toronto Press, 1965.

Nutter, G.W. 1951. *The extent of enterprise monopoly in the United States, 1899–1939*. Chicago: University of Chicago Press.

Penny Cyclopedia of the Society for the Diffusion of Useful Knowledge. 1839.

Schumpeter, J.A. 1942. *Capitalism, socialism and democracy*. New York: Harper & Bros.

Senior, N.W. 1836. *Political economy*. London: W. Clowes.

Shepherd, W.G. 1982. Causes of increased competition in the US economy, 1939–1980. *Review of Economics and Statistics* 64: 613–626.

Smith, A. 1776. *The wealth of nations*. Glasgow ed. Oxford: Oxford University Press. 1976.
Stigler, G.J. 1957. Perfect competition, historically contemplated. *Journal of Political Economy* 65: 1–17.
Stigler, G.J., and R. Sherwin. 1985. The extent of the market. *Journal of Law and Economics* 28: 555–585.
Thornton, W.T. 1869. *On labour*. London: Macmillan.

Competition and Efficiency

John C. Panzar

The association between economic efficiency and competition goes back at least as far as Adam Smith's 'invisible hand' metaphor. Indeed, a goodly portion of the vast body of subsequent work in value theory has dealt with the normative issues arising from the workings of the competitive economy. Thus any short essay on the topic must be somewhat idiosyncratic, focusing upon the points which are of greatest interest to the author. Therefore, I shall limit my attention to the properties of the (static, partial equilibrium) *economic model* of perfect competition and how its use has recently been extended to add to our understanding of a larger range of real world markets. I must leave to others the tasks of sorting out the importance of competition in, for example, the Schumpeterian process of 'creative destruction', the aggregation and transmission of society's stock of information, or the evolutionary progress of technological advance. Fortunately for my purposes, the historical development of the competitive model has been thoroughly analysed by Stigler (1957). The formulation of the model, as we know it today, was completed in the work of Knight (1921). It is interesting to note that the last refinement to be added was the free mobility of resources across industries: i.e. the entry and exist of firms. In his insightful concluding section, Stigler points out that competition can flourish *within* a market without this last ingredient. (Consider an agricultural market with Ricardian rents.) He suggested that the term 'market competition' be used to describe such situations, and that the term 'industrial competition' be applied when mobility across industries is present. The work that I shall discuss deals with the converse possibility: perfectly contestable markets, situations in which competition may not necessarily exist within a particular market, but firms (and resources) are assumed to be perfectly mobile across industries.

The role of entry and exit in assuring the equalization of returns across markets is not logically limited those cases in which it is technologically feasible for the market to be populated by a large number of firms, each capable of achieving an efficient scale of operation. It may be expected that the lure of profits might serve to make relevant certain aspects of competitive theory even under conditions of 'natural monopoly'. The most striking practical illustration of this point was the recent deregulation of airlines in the United States. This took place, in part, because the free mobility of resources (aircraft) across markets led policy makers to believe that satisfactory economic performance could be achieved without the stultifying effects of economic regulation. This, despite the fact that most city-pair airline markets are natural monopolies and none can be expected to support the large numbers of firms required by the perfectly competitive model. Thus the need to extend at least part of the competitive paradigm to incorporate such cases had become apparent.

In a classic article, Demsetz (1968) set forth one way to break the commonly perceived link between monopoly provision of certain increasing returns services and monopoly conduct on the part of the firm providing the service at any point in time. By pointing out that the impossibility of competition within the market need not preclude effective competition *for the market*, Demsetz raised a fundamental challenge to the conventional wisdom that the only effective ways to deal with a technological natural monopoly were through economic regulation or public enterprise.

Demsetz chose to elaborate this idea in the context of a franchise bidding scheme, in which the franchise was to be awarded, not to the firm offering the greatest lump sum payment to the municipal coffers, but to the firm offering to

serve the market at the lowest price. Subsequent authors have criticized this as a policy proposal, focusing on the problems raised by considerations of sunk costs and incomplete contracts, from which Demsetz explicitly sought to abstract. However, there is another sense in which the franchise bidding example may have been an unfortunate expository choice. Because it introduced a new institution between the firm and the market – the franchise auctioneer – this illustration may have obscured the link between the analysis of competition for the market and the earlier notion of the role of free entry and exit in ensuring effective industrial competition. Furthermore, Demsetz's simple bidding scheme cannot handle the realistic cases in which the monopolist produces two or more technologically related services.

The theory of contestable markets developed by Baumol et al. (1982) is most usefully viewed as an attempt to extend the neoclassical (partial equilibrium) theory of long-run competitive equilibrium to the case of increasing returns to scale. In so doing, they developed a model which achieved the Demsetz solution to the monopoly problem as the result of a market equilibrium process. This extension was accomplished by emphasizing the role played by potential entry in characterizing the role defining properties of long-run competitive equilibrium. To see this reinterpretation most clearly, the following definitions are necessary:

Definition 1 A Feasible Industry Configuration (FIC) is a collection of firms, $i = 1, \ldots, m$ output vectors for each, $y^1, \ldots, y^m$; and a market price vector p such that each firm earns non-negative profits and the total quantity supplied equals the quantity demanded; i.e. $py^i - C(y^i) \geq 0$, for all $i = 1, \ldots, m$ and $\sum y^i = D(p)$, where C is the (multiproduct) minimum cost function and D the market demand function.

Feasibility surely reflects the minimal conditions one would expect to prevail in long-run industry equilibrium in a private enterprise economy: All firms must earn non-negative profits and the total quantity supplied by firms equals the amount demanded by consumers at the market price. While feasibility requires financial viability, it does not preclude the positive profits which may attract entry. Therefore, the neoclassical notion of long-run competitive equilibrium must encompass some additional restrictions. More specifically,

Definition 2 A long-run competitive equilibrium is any FIC which also has the property that $py - c(y) \leq 0$ for all y.

While this characterization of long-run competitive equilibrium may be unfamiliar, it is equivalent to the standard notion of price taking firms earning zero economic profits by equating marginal cost to price. (To see this, note that since profits are nonpositive for all output levels, the fact that $py^1 - C(y^i) \geq 0$ means that output level y^i maximizes the ith firm's profits. This, in turn, implies that $MC(y^i) = p$ if firm i is producing.)

Characterizing competitive equilibrium via Definitions 1 and 2 has the advantage of focusing attention on the role played by potential entry. The strictures of Definition 2 can be interpreted to mean that the firms in an industry in long-run competitive equilibrium act as if they were policed by potential entrants prepared to enter the market in pursuit of any profit opportunity calculated at current market prices. While this lack of attention to the possibility of retailatory price responses by rivals reflects the noncooperative spirit of the competitive paradigm, it ignores the response of consumers to a change in the market price. Therefore it is useful to consider making potential entrants 'less optimistic' in the following sense:

Definition 3 A Sustainable Industry Configuration (SIC) is any FIC which also satisfies the condition that $p^e y^e - C(y^e) \leq 0$ for all $p^e \leq p$ and $y^e \leq D(p^e)$.

Thus firms in a SIC behave as if the market were policed by potential entrants that calculate the profitability of entry under the assumption that incumbent firms' prices remained unchanged, but that do take account of the reality that consumers can be induced to purchase a larger quantity only at a lower price. Put another way, an FIC is also an SIC when no potential

entrant can anticipate earning a positive profit by quoting a price at or below that prevailing in the market and serving all or a part of the resulting demand. The following semantic clarification completes the characterization of a contestable market:

Definition 4 A perfectly contestable market is one in which perfectly free entry and exit ensure that the only possible long-run equilibria are SICs.

An immediate implication of Definitions 1, 2, 3, and 4 is:

Proposition 1 Any long-run competitive equilibrium is an SIC, but not conversely. Thus all perfectly competitive markets are perfectly contestable, but not all perfectly contestable markets are perfectly competitive.

The proof follows from the fact that the conditions which an FIC must satisfy in order to be a long-run competitive equilibrium are stronger than those required of an SIC. Thus a long-run competitive equilibrium is, by construction, an SIC. To see that the converse is not true, consider the case in which the average costs of production fall throughout the relevant range; i.e. at least as far as the intersection of the average cost curve and the market demand curve. The point of intersection, the Demsetz outcome, characterizes a sustainable industry configuration, since profits are non-positive and no point on or below the demand curve can yield non-negative profits at a lower price. However this outcome is clearly not a long-run competitive equilibrium because price is equal to average cost which, by hypothesis, is strictly greater than marginal cost. The above demonstration points out the fact that the concept of contestable markets can be applied beyond the large-numbers case of perfect competition. However it also raises questions about the efficiency properties of such markets. The fact that equilibrium may involve a price greater than marginal cost means that the First Best optimality properties of the competitive model need no longer apply. What efficiency properties, then, *can* be associated with equilibria in contestable markets? In the case of single product markets it is intuitively clear (and straightforward to prove) that, when they exist, sustainable industry configurations are solutions to the Second Best optimization problem: maximize welfare (as measured, for example, by the sum of producers' and consumers' surpluses) subject to the constraint that firms earn non-negative profits. Clearly, when increasing returns to scale render marginal cost pricing unprofitable, the best that can be done, in the absence of lump sum transfers and discriminatory or non-linear pricing, is to set price equal to average cost.

However, even this level of performance can no longer be guaranteeed once one moves to the realistic realm of multiple products. For example, a monopolist producing two or more products can, in general, find an infinite number of price combinations which will yield it exactly zero economic profits. Some of these prices and resulting market demand quantities may represent SICs. Call this set *P*. If the underlying cost and demand functions are sufficiently well-behaved, there will exist a unique constrained welfare maximizing price vector p^*. The most desirable efficiency result would be for the set P to consist of the single element p^*. Unfortunately, it is easy to construct examples in which P does not contain p^*, as well as cases in which P is empty. What efficiency properties does this generalized process of industrial competition possess when extended beyond the realm of perfect competition? The results that pertain generally lie entirely on the cost side.

Proposition 2 In any SIC, the industry's output is divided among the firms in a way that minimizes total industry costs.

The proof is by contradiction. Consider an initial SIC composed of m firms producing output vectors, $y^1, \ldots, y^m$, at market prices p. Suppose, contrary to hypothesis, that there exists an alternative group of k firms with output vectors, $z^1, \ldots, z^k$, that could produce the current industry output at a lower total cost. That is $\sum_j z^j = D(p) = \sum_i y^i$, but $\sum_j C(z^j) < \sum_i C(y^i)$. Then the new group, in total, would earn positive economic profits at the initial price p. This is true because, by hypothesis, total revenues would be equal, but total costs would be lower for the alternative group, while the initial

group of firms must have been earning non-negative profits. Therefore at least one firm, say firm j, in the alternative group would anticipate earning strictly positive profits at the price vector p. But then there exists an entry plan $p^e = p \leq p$ and $y^e = z^j \leq D(p^e)$ such that $p^e y^e - C(y^e) > 0$, which contradicts the hypothesis that the initial group of firms constituted a SIC.

Additional efficiency results for contestable markets are presented in chapter 11 of Baumol, Panzar and Willig. Here, I shall mention specifically a class of results which are relevant only in the multiproduct context. One implication of the fact that equilibrium in a contestable market presents no profit opportunities for potential entrants is that no subset of services of a multiproduct enterprise can generate revenues in excess of the cost of providing them alone. Thus, equilibrium in perfectly contestable markets cannot involve one gorup of services subsidizing another. Whether or not this property is a desirable efficiency result is unclear. Consider, for example, a situation in which a monopoly firm uses common facilities to produce two services, one of which has a very elastic demand curve while that of the other is very inelastic. Maximizing total surplus subject to a break-even constraint leads to the well-known inverse elasticity rule: the markup of price over marginal cost is greater for services whose demand is least elastic. However, this pricing policy may easily lead to revenues from the inelastic service in excess of the cost of providing it alone. Such an outcome would not be an SIC and could not persist in a perfectly contestable market. Thus while the mobility of firms and resources can, even without the presence of market competition, ensure productive efficiency, it cannot in general guarantee that an optimal relationship of output prices in a multiproduct industry will prevail outside the perfectly competitive realm.

See Also

- ▶ Contestable Markets
- ▶ Increasing Returns to Scale
- ▶ Monopoly
- ▶ Natural Monopoly

References

Baumol, W., J. Panzar, and R. Willig. 1982. *Contestable markets and the theory of industry structure*. New York: Harcourt Brace Jovanovich.

Demsetz, H. 1968. Why regulate utilities? *Journal of Law and Economics* 11: 55–65.

Knight, F. 1921. *Risk, uncertainty, and profit*. Chicago: University of Chicago Press.

Stigler, G. 1957. Perfect competition, historically contemplated. *Journal of Political Economy* 65: 1–17.

Competition and Selection

Sidney G. Winter

Abstract

The claim that a business firm must maximize profit if it is to survive serves as an informal statement of the common conclusion of a class of theorems characterizing explicit models of economic selection processes. Such models, by making explicit the strong assumptions needed to generate this sort of result, are the basis for a critique of standard economic theory which relies on competitive equilibrium. Models of Schumpeterian competition, emphasizing the centrality of innovation, plainly provide a much better description of the world we live in than do models of static equilibrium.

Keywords

Adjustment, dynamic vs static; Alchian, A.; Behavioural change; Comparative statics; Competition and selection; Competitive equilibrium; Constant returns to scale; Decreasing returns to scale; Diminishing returns; Entrepreneurial rents; Entrepreneurship; Evolutionary economics; Fixed factors; Friedman, M.; Increasing returns to scale; Innovation; Latent productivity; Market power; Mimicry theorems; Natural selection; Neoclassical growth theory; Patents; Present value; Profit maximization; Research and development; Rules of

behaviour; Satisficing; Schumpeter, J.; Schumpeterian competition; Selection equilibrium; Simon, H.; Static equilibrium; Survival of the fittest; Technological opportunity; Winter, S. G.

JEL Classifications
B0

Under competitive conditions, a business firm must maximize profit if it is to survive – or so it is often claimed. This purported analogue of biological natural selection has had substantial influence in economic thinking, and the proposition remains influential today. In general, its role has been to serve as an informal auxiliary defence, or crutch, for standard theoretical approaches based on optimization and equilibrium. It appeared explicitly in this role in a provocative passage in Milton Friedman's famous essay on methodology (Friedman 1953, ch. 1), and it seems that many economists are familiar with it in this context only.

There is, however, an alternative role that the proposition can and does play. It serves as an informal statement of the common conclusion of a class of theorems characterizing explicit models of economic selection processes. A model in this class posits, first, a range of possible behaviours for the firm. This range must obviously extend beyond the realm of profit maximization if the conclusion of the argument is to be non-trivial, and it must include behaviour that is appropriately termed 'profit maximizing' if the conclusion is to be logically attainable at all. The model must also characterize a particular dynamic process that in some way captures the general idea that profitable firms tend to survive and grow, while unprofitable ones tend to decline and fail. A stationary position of such a process is a 'selection equilibrium'.

Models of this type occupy an important but non-central position in evolutionary economic theory (Nelson and Winter 1982). They establish that the equilibria of standard competitive theory can indeed be 'mimicked' (in several different senses) by the equilibria of selection models. More importantly, by making explicit the strong assumptions that apparently are required to generate this sort of result, they are the basis for a critique of its generality and an appraisal of the strength of the crutch on which standard theory leans. They also provide a helpful entry-way to the much broader class of evolutionary models in which mimicry results fail to hold. This entry-way has the convenient feature that the return path to standard theory is well marked; the sense in which evolutionary theory subsumes portions of standard theory becomes clear.

The concept of competition need not, of course, be considered only in the context of perfectly competitive equilibrium. In a broader sense of the term, any nontrivial selection model in which the 'fit' prosper and the 'unfit' do not is a model of a 'competitive' process. The process need not have a static equilibrium, or any equilibrium, and it may easily lead to results that are clearly non-competitive by the standards of industrial organization economics.

The remainder of this essay first considers in more detail the theoretical links between selection processes and competitive equilibrium outcomes. It then examines a more interesting and less well-explored area that involves selection and, in a broad sense, competition; Schumpeterian competition.

Competitive Equilibrium as a Selection Outcome

The intention here is to describe the heuristic basis of existing examples of this type of theorem, or, alternatively, to describe the basic recipe from which an obviously large class of broadly similar results could be produced. There may be other basic recipes, as yet unknown. There certainly are ways to ignore individual instructions of the recipe and yet preserve the result, though at the cost of delicately contrived adjustments in other assumptions.

(To avoid confusion, it should be noted at the outset that the word 'equilibrium' is used in two different senses in this discussion, the 'no incentives to change behaviour' sense employed in economic theory and the 'stationary position of a dynamic process' sense that is common outside of

economics. The point of the discussion is, in fact, to relate these two equilibrium ideas in a particular way.)

(1) Constant returns to scale must prevail in the specific sense that the supply and demand functions of an individual firm at any particular time are expressible as the scale (or 'capacity') of that firm at that time multiplied by functions depending on prices, but not directly on scale or time. Increasing returns to scale must be excluded for familiar reasons. Decreasing returns must be excluded because they will in general give rise to equilibrium 'entrepreneurial rents' which could be partially dissipated by departures from maximization without threatening the survival of the firm. Thus, for example, the U-shaped long run average cost curve of textbook competitive theory does not provide a context in which selection necessarily mimics standard theory if competitive equilibrium would require some firms to be on the upward sloping portion of the curve.

(2) Firms must increase scale when profitable and decrease scale (or go out of business entirely) when unprofitable. Alternatively, profitability of a particular firm must lead to entry by perfect imitators of that firm's actions. In the absence of such assumptions, it is plain that there will in general be equilibria with non-zero profit levels, which under assumption (1) cannot mimic the competitive result. While the 'decline or fail' assumption is a plausible reflection of long-run breakeven constraints characteristic of actual capitalist institutions, no such realistic force attaches to the requirement that profitability lead to expansion. If firms do not pursue profits in the long-run sense of expanding in response to positive profitability, stationary positions may involve positive profits. Such stationary positions fail to mimic competitive equilibria for that reason alone (given constant returns), but they also introduce once again the possibility that the short-run behavioural responses of surviving firms may dissipate some of the positive profit that is potentially achievable at selection equilibrium scale.

In standard theory, expansion in response to profitability may be seen as an aspect of the firm's profit-seeking on the assumption that it regards prices as unaffected by its capacity decisions. In turn, this ordinarily requires that the firm in question be but one of an indeterminately large number of firms that all have access to the same technological and organizational possibilities.

While the assumption that firms have identical production sets and behavioural rules is common and appears inoffensive in orthodox theorizing, it is very much at odds with evolutionary theory. The orthodox view comes down to the assertion that all productive knowledge is freely available to one and all – perhaps it is all in the public library. By contrast, evolutionary theory emphasizes the role of firms as highly individualized repositories of productive knowledge, not all of which is articulable. From the evolutionary perspective, the fact that mimicry theorems rely on assumptions of unimpaired access to a public knowledge pool is by itself sufficient to make it clear that the selection argument can provide only a weak and shaky crutch for standard competitive theory.

(3) A firm that is breaking even with a positive output at prevailing prices must not alter its behaviour; a potential entrant that would only break even at prevailing prices must not enter. This assumption is needed to assure that the competitive equilibrium position is in fact a stationary position of the selection process.

Models of natural selection in biology do not typically involve this sort of assumption, but neither do they conclude that only the fittest genotypes survive – the biological analogue of the proposition discussed here. Rather, they show how constant gene frequencies come to prevail as the selection forces that tend to eliminate diversity come into balance with mutation forces that constantly renew it. A strictly analogous treatment of economic selection would be much more appealing than the sort of result discussed here.

It would admit that occasional disruptions may arise from random behavioural change, or from over-optimistic entrants. Thus, potentially at least, it could better serve the purpose of establishing the point that the results of standard competitive theory are in some sense robust with respect to its behavioural assumptions. Unfortunately, standard theory offers no clue as to what this sense might be. It is plain that the adjustment processes of the system are centrally involved, and there is no behaviourally plausible theory of adjustment that is the dynamic counterpart in the disciplinary paradigm of static competitive equilibrium theory.

Within the limits defined by the requirement for a strictly static competitive outcome, the most plausible approach combines the idea of characterizing the firms in the selection process by their 'rules of behaviour' – an idea advanced in a seminal paper by Armen Alchian (1950) – with Herbert Simon's idea of satisficing (1955). In the simplest version, each firm simply adheres unswervingly to its own deterministic behavioural rule (or 'routine', in the language of Nelson and Winter 1982). Such a rule subsumes or implies the firm's supply and demand functions, and given the conditions set forth in (1) and (2) above, a constant environment evokes a constant response. Satisficing may be introduced as a complication of this picture by an assumption that a firm that sustains losses over a period of time will search for a better behavioural rule; this adds behavioural plausibility to the adjustment process but does not introduce the possibility that random rule change might disrupt an otherwise stationary competitive equilibrium position.

(4) The final requirement can be succinctly but inadequately stated as 'some firms must actually be profit maximizers'. Although this formulation does adequately cover some simple cases, it does not suggest the depth and subtlety of the issues involved.

Two points deserve particular emphasis here. The first is the distinction between profit maximizing *rules of behaviour* (functions) and profit maximizing *actions.* In general, a selection equilibrium that mimics a particular competitive equilibrium must clearly be one in which some firms take actions that are profit maximizing in that competitive equilibrium, and in this sense are profit maximizers. But this observation does not imply that the survivors in the selection equilibrium possess maximizing *rules*, and in general it is not necessary that survivors be maximizers in this stronger sense. (Proof: Consider a competitive equilibrium with constant returns to scale. Restrict the firms' supply and demand functions to be constant up to a scale factor at the values taken in the given equilibrium. Embed this static equilibrium in a dynamic adjustment system in which firms' scales of output respond to profitability in accordance with assumption (2). Then the given competitive equilibrium becomes a selection equilibrium – since the only techniques in use make zero profit – but the firms are not profit maximizers in the stronger sense.)

The second point extends the first. The notion of profit maximizing behavioural rules itself rests on the conceptual foundation of a production set or function that is regarded as a given. In evolutionary theory, however, it is the rules themselves that are regarded as data and as logically antecedent to the values (actions) they yield in particular environments. Thus, in this context, a problem arises in interpreting the basic idea of a selection equilibrium mimicking a standard competitive one: there is no obvious set of 'possibilities' to which one should have reference.

The most helpful approach here emphasizes internal consistency. Assumptions about the structure of what is 'possible' can be invoked without the additional assumption that there is a given set of possibilities – for example, additivity and divisibility may be assumed without implying that the set of techniques to which these axioms apply is a given datum of the system. Such an approach provides a basis for discussing whether a particular selection equilibrium is legitimately *interpretable* as a competitive equilibrium given the other assumptions in force. Along this path one can explore a rich variety of selection equilibrium situations that may be thought of as competitive equilibria. Precisely because the variety is so rich, to know only that an outcome is interpretable in this fashion is to know very little about it.

In the light of formal analysis of selection models of the sort described above, how strong is the crutch that selection provides to standard theory? For many analytical purposes, it is a crucial weakness that the crutch relates only to equilibrium actions and not to behavioural rules; it is from the knowledge that the rules are maximizing that the results of comparative statics derive. A selection system disturbed by a parameter change from a 'mimicking' equilibrium does not necessarily go to a new 'mimicking' equilibrium, let alone to one that is consistent, in standard theoretical terms, with the information revealed in the original equilibrium. More fundamentally, selection considerations cannot compensate for the inadequacies of standard theory that arise from the basic assumption that production possibilities are given data of the system.

Schumpeterian Competition

In two great works and in many other writings, Joseph Schumpeter proclaimed the central importance of innovative activity in the development of capitalism. His early book, *The Theory of Economic Development,* focused on the role and contribution of the individual entrepreneur. From today's perspective the work remains enormously insightful and provocative but may seem dated; the image of the late 19th-century captains of industry lurks implicitly in the abstract account of the entrepreneur. The late work, *Capitalism, Socialism and Democracy*, is likewise insightful, provocative and a bit anachronistic. In this case, the anachronism derives from the predictions of a future in which the innovative process is bureaucratized, the role of the individual entrepreneur is fully usurped by large organizations, and the sociopolitical foundations of capitalism are thereby undercut. Present reality does not correspond closely to Schumpeter's predictions, and it seems increasingly clear that he greatly underestimated the seriousness of the incentive problems that arise within large organizations, whether capitalist corporations or socialist states.

Substantial literatures have accumulated around a number of specific issues, hypotheses and predictions put forward in Schumpeter's various writings. Regardless of the verdicts ultimately rendered on particular points, everyday observation repeatedly confirms the appropriateness of his emphasis on the centrality of innovation in contemporary capitalism. It confirms, likewise, the inappropriateness of the continuing tendency of the economics discipline to sequester topics related to technological change in subsectors of various specialized fields, remote from the theoretical core.

The purpose of the present discussion is to assess the relationships of selection and competition from a Schumpeterian viewpoint, that is, to extend the discussion above by considering what difference it makes if firms are engaged in inventing, discovering and exploring new ways of doing things. Plainly, one difference it makes is that 'competition' must now be understood in the broad sense that admits a number of additional dimensions to the competitive process, along with price-guided output determination. In particular, costly efforts to innovate, to imitate the innovations of others, and to appropriate the gains from innovation are added to the firm's competitive repertoire.

Selection now operates at two related levels. The organizational routines governing the use made of existing products and processes in every firm interact through the market place, and the market distributes rewards and punishments to the contenders. These same rewards and punishments are also entries on the market's scorecard for the higher level routines from which new products and processes derive – routines involving, for example, expenditure levels on innovative and imitative R&D efforts. Over the longer term, selection forces favour the firms that achieve a favourable balance between the rents captured from successive rounds of innovation and the costs of the R&D efforts that yield these innovations.

In formal models constructed along these lines, it is easy to see how various extreme cases turn out. One class of cases formalizes the cautionary tale told by Schumpeter (1950, p. 105), in which competition that is 'perfect – and perfectly prompt' makes the innovative role non-viable.

Sufficiently high costs of innovation and low costs of imitation (including costs of surmounting any institutional barriers such as patents) will lead to the eventual suppression of all firms that continue to attempt innovation, and the system will settle into a static equilibrium. (The character of this equilibrium may, however, depend on initial conditions and on random events along the evolutionary path; the production set ultimately arrived at is an endogenous feature of the process.) One can also construct model examples to illustrate the cautionary message 'innovate or die', the principal requirement being simply a reversal of the cost conditions stated above.

With the exception of some extreme or highly simplified cases, models of Schumpeterian competition describe complex stochastic processes that are not easily explored with analytical methods. Of course, the activity of writing down a specific formal model is often informative by itself in the sense that it illuminates basic conceptual issues and poses key questions about how complex features of economic reality can usefully be approximated by a model. Some additional insight can then be obtained using simulation methods to explore specific cases (Nelson and Winter 1982, Part V; Winter 1984). One of the most significant benefits from simulation is the occasional discovery of mechanisms at work that are retrospectively 'obvious' and general features of the model.

The discussion that follows pulls together a number of these different sorts of insights, emphasizing in particular some issues that do not arise in the related theoretical literature that explores various Schumpeterian themes using neoclassical techniques (For the most part these neoclassical studies explore stylized situations involving a single possible innovation, and thus do not address issues relating to the cumulative consequences of dynamic Schumpeterian competition. See Kamien and Schwartz (1981) and Dasgupta (1985) for references and perspectives on this literature.)

A fundamental constituent of any dynamic model of Schumpeterian competition is a model of technological opportunity. Such a model establishes the linkage between the resources that model firms apply to innovative effort and their innovative achievements. The long run behaviour of the model as a whole depends critically on the answers provided for a set of key questions relating to technological opportunity. Does the individual firm face diminishing returns in innovative achievement as it applies additional resources over a short period of time? If so, from what ' fixed factors' does the diminishing returns effect arise, and to what extent are these factors subject to change over time either by the firm's own efforts or by other mechanisms? Are selection forces to be studied in a context in which technological opportunity presents more or less the 'same problem' for R&D policy over an extended period, or is the evolutionary sorting out of different policies for the firm a process that proceeds concurrently with historical change in the criteria that govern the sorting?

Technological opportunity is said to be *constant* if R&D activity amounts to a search of an unchanging set of possibilities – in effect, there is a meta-production set or meta-production function that describes what is ultimately possible. *Increasing* technological opportunity means that possibilities are being expanded over time by causal factors exogenous to the R&D efforts in question – implying that, given a level of technological achievement and a level of R&D effort, the effort will be more productive of innovative results if applied later. With constant technological opportunity, returns to R&D effort must eventually be decreasing, approaching zero near the boundary of the fixed set of possibilities.

It is all too obvious that it may be very difficult to develop an empirical basis for modelling technological opportunity in an applied analysis of a particular firm, industry or national economy. There is no easy escape from the conundrum that observed innovative performance reflects both opportunity and endogenously determined effort, not to mention the fact that neither performance nor effort is itself easily measured or the even more basic question of whether analysis of the past can illuminate the future. These difficulties in operationalizing the concept of technological opportunity do not, unfortunately, in any way diminish its critical role in Schumpeterian competition.

The evolutionary analysis of Schumpeterian competition has not, thus far, produced any counterpart for the sorts of mimicry theorems that can be proved for static equilibria. That is, there is no model in which it can be shown that selection forces, alone or in conjunction with adaptive behavioural rules, drive the system asymptotically to a path on which surviving firms might be said to have solved the remaining portion of the dynamic optimization problem with which the model situation confronts them – except in the cases where the asymptotic situation is a static equilibrium with zero R&D. The list of identified obstacles to a non-trivial positive result is sufficiently long, and the obstacles are sufficiently formidable, so as to constitute something akin to an impossibility theorem. It seems extremely unlikely that a positive result can be established within the confines of an evolutionary approach – that is, without endowing the model firms with a great deal of correct information about the structure of the total system in which they are embedded.

The most formidable obstacle of all derives from the direct clash between the future-oriented character of a dynamic optimization and the fact that selection and adaptation processes reflect the experience of the past. If firms cannot 'see' the path that technological opportunity will follow in the future, if their decisions can only reflect past experience and inferences drawn therefrom, then in general they cannot position themselves optimally for the future. They might conceivably do so if the development of technological opportunity were simple enough to validate simple inference schemes. Such simplicity does not seem descriptively plausible; who is to say that it is implausible that in a particular case technological opportunity might be constant, or exponentially increasing, or following a logistic, or some stochastic variant of any of these? And without some restriction on the structural possibilities, how are model firms to make inferences to guide their R&D policies?

This obstacle is not featured prominently in the simulations reported by Nelson and Winter, which are largely confined to very tame and stylized technological regimes in which opportunity is summarized by a single exponentially increasing variable, called 'latent productivity'. Such an environment, reminiscent in some ways of neoclassical growth theory, seems at first glance to be a promising one for the derivation of a balanced growth outcome in which actual and latent productivity are rising at the same rate, the problem facing the firms is in a sense constant, and selection and adaptation might bring surviving firm R&D policies to optimal values.

In fact, such a result remains remote even under the very strong assumption just described. Demand conditions for the product of the industry (or the economy) affect the long run dynamics, and in this area also assumptions must be delicately contrived to avoid excluding a balanced growth outcome. For example, consider an industry model with constant demand in which demand is (plausibly) less than unit elastic at low prices. Then, cost reduction continued indefinitely would drive sales revenue to zero. Zero sales revenue will not cover the cost of continuing advance. What is involved here is a reflection of the basic economics of information; costs of discovery are independent of the size of the realm application, and on the assumption stated the economic significance of that realm is dwindling to nothing. The implication is that demand conditions may check progress even if technological opportunity is continually expanding. Indeed, this may well be the pattern that is typically realistic for any narrowly defined sector.

This difficulty too can be dispatched by an appropriately chosen assumption. Beyond it lie some further problems. A model that acknowledges the partially stochastic nature of innovative success will display gradually increasing concentration (Phillips 1971), unless some opposing tendency is present. A good candidate for an opposing tendency is the actual exercise of market power that has been acquired by chance (Nelson and Winter 1982, ch. 13). But this market power can, presumably, also shelter various departures from present value maximization, including departures from dynamically optimal R&D policy.

To reiterate, the quest for mimicry theorems in the context of Schumpeterian competition seems foredoomed to failure. Since models of

Schumpeterian competition plainly provide a much better description of the world we live in than do models of static equilibrium, the overall conclusion with regard to the strength of the selection crutch is distinctly more negative than the conclusion for static models alone. Assumptions that firms maximize profit or present value will have to stand on their own, at least until somebody invents a better crutch for them. In the meantime, it will continue to be the case that predictions based on these assumptions are sometimes sound and sometimes silly, and standard theory does not offer a means of discriminating between the cases. More direct attention should be paid to the mechanisms of selection, adaptation and learning, which among them probably account for as much sense as economists have actually observed in economic reality, and also leave room for a lot of readily observable nonsense.

Bibliography

Alchian, A.A. 1950. Uncertainty, evolution and economic theory. *Journal of Political Economy* 58: 211–221.

Dasgupta, P. 1985. The theory of technological competition. In *New developments in the analysis of market structure*, ed. J. Stiglitz and G.F. Mathewson. Cambridge, MA: MIT Press.

Friedman, M. 1953. *Essays in positive economics*. Chicago: University of Chicago Press.

Kamien, M., and N. Schwartz. 1981. *Market structure and innovation*. Cambridge: Cambridge University Press.

Nelson, R., and S. Winter. 1982. *An evolutionary theory of economic change*. Cambridge, MA: The Belknap Press of Harvard University Press.

Phillips, A. 1971. *Technology and market structure: A Study of the Aircraft Industry.* Lexington: D.C. Heath.

Schumpeter, J.A. 1912. *The theory of economic development.* Trans. Redvers Opie. Cambridge, MA: Harvard University Press, 1934.

Schumpeter, J.A. 1950. *Capitalism, socialism and democracy*, 3rd ed. New York: Harper.

Simon, H. 1955. A behavioral model of rational choice. *Quarterly Journal of Economics* 69: 99–118.

Winter, S.G. 1964. Economic 'natural selection' and the theory of the firm. *Yale Economic Essays* 4(1): 225–272.

Winter, S.G. 1971. Satisficing, selection and the innovating remnant. *Quarterly Journal of Economics* 85: 237–261.

Winter, S.G. 1984. Schumpeterian competition in alternative technological regimes. *Journal of Economic Behaviour and Organization* 5(3–4): 287–320.

Competition in International Trade

D. K. Stout

In international markets the equivalent of the owned assets and the skills of individuals and firms in internal markets is the difference in relative national resource endowments. In elementary theory, the effects of international competition are defined by comparison with an initial state of no trade. The typical question addressed is 'What will happen if trade is opened up?' On this view, international competition 'causes' trade because of the differences there would be, in the absence of trade, in the relative costs of production of pairs of goods in two countries. And the effect of international competition, and the ensuing trade, is to have ironed out some of these differences and to have increased in the process the aggregate equilibrium output of each good.

In each country, initially, the cheaper good is the one whose production technology calls for a lot of the input that is relatively plentiful there. When trade is opened up, the producers of the dearer commodity in each country face competition from the country better suited to produce it. Through international competition, differences in the relative scarcity (and cost) of labour, land and capital and consequent differences in the ratio of the price of, say, food to the price of, say, machinery are lessened. With the growth of multinational companies (MNCs), choices are made between competing by offering products (i.e. by exporting) and by licensing domestic producers overseas or setting up subsidiary companies to produce and market on the spot. These choices depend upon the importance of economies of scale in production; transport costs; the costs of transferring capital, technology and management; and upon the restrictions imposed by governments upon the free movement of goods: that is, tariffs, import quotas and other barriers.

If all industries were subject to diminishing returns and perfect competition ruled, with homogeneous products, differences in the factor

endowments of two countries starting to compete would lead to specialization up to the point where, at the margin, the relative costs of each pair of goods were the same in each country. In fact, with increasing returns to scale, equal relative marginal costs are no guarantee of equilibrium. Competitive advantage may accrue to an industry in the country where home demand expands and is protected for a time, so that relative costs fall as output increases. This so-called 'infant industry' case is still the most striking illustration of the inadequacy of simple static theorems about the effects of international competition upon trade flows.

The account of trade flows due to Heckscher and Ohlin (namely that it depends on differing input requirements for different goods and differing input endowments in different countries) remains the most important observation about international competitive advantage.

This model of specialization through trade can be modified to cope quite well with such market imperfections as transport and other transaction costs, limited information, a degree of imperfect competition in factor markets and limited trade barriers. It becomes seriously strained however when one tries to use it to explain the results of international competition in the world of many traders and many commodities. The number of separate 'factors of production' has to be increased to equal the number of traded commodities and defined to suit the observed trade flows. Competitive advantage became something 'revealed', rather than predictable by taking a census of resources. Cheap US skilled labour might be redefined as capital; but since physical capital was also cheap in the US the paradox which Leontief observed – that the US tended to import capital-intensive goods and export labour-intensive – could not be resolved this way.

On another view, out of which grew an alternative and more robust model of international competition, high technology was regarded as complementary to (skilled) labour explaining the bias of US trade. But a theory which relied first on separating factors of production but then required them to be lumped together is not a good theory. Furthermore, competitive advantages were found to be continuously changing in ways that could not be predicted by counting heads or hectares.

In the 1950s and 1960s, when industrial international competition became much more open and many countries were newly industrializing, trade both ways in different variants of the same products grew prodigiously. It has become clear that international competition drives intraindustry trade in imperfectly competitive differentiated markets; that in very many markets today's net exporter is tomorrow's importer; and that international competition involves the international transfer of capital and knowhow as well as of goods.

Something is needed then to supplement the explanation of broad directions of trade advantage in terms of fixed differences in factor endowments and given technology. There has arisen a less determinate, more detailed account of disequilibrium change in trade flows under conditions of imperfect competition. This account makes it clear how international competition leads to two-way flows of trade within the same industry. Where there are economies of scale in each of eight plants, four in each country, producing four variants of one product, a likely outcome is that four of the eight will survive, perhaps two in each country, and in each country the loyal customers of each type will import or buy at home accordingly. *Net* trade flows might be zero but gross trade might be half of total consumption. Or two countries, one rich and one poor, might have developed at home (through domestic competition) lower relative costs in the 'superior' variant in the richer country and in the 'down-market' variant in the poorer country. As trade becomes easier, with cheaper transport and communications or lower tariff barriers, the poor country exports its variant to the poorer consumers in the richer country and vice versa. As incomes grow in both countries, demand in both will shift more and more towards the rich country's variant, and the poorer country loses overall market share.

Where the pressures of competition are leading to successive shifts in state-of-the-art technology in a discontinuous way, competitive advantage may move from the technological leader nation to the fast followers as they acquire this

technology; and back to the leader again with the next major breakthrough, as has been observed in industries as different as radio receivers and textiles.

These models of changing comparative advantage owe much to Joseph Schumpeter. They focus upon the temporary monopoly advantages that accrue to the innovator, extended sometimes by patent protection and economies of scale.

As the technology gap is closed by imitators competing away the monopoly profits, the technological leader nation, like Schumpeter's individual innovator, jumps to new dry steppingstones in today's 'sunrise' industries. The technology gap is moved but kept open.

Whenever this evolution involves new products, it becomes closely linked to a now well-established view of the development of product markets: namely, that products move through a life cycle of youthful growth as comparative luxuries; into explosive growth as they are mass-produced and become available to poorer consumers; into maturity, stagnation and sometimes death as they are superseded by a more efficient or desirable way of satisfying the same underlying needs. As products move through their life cycles, international competitive advantage will tend to shift from developed technological leader economies to newer or poorer industrializing economies, particularly those with large domestic markets allowing home-grown scale economies.

Technology gap and product life-cycle theories do not explain why some economies lead and others follow. The underlying point is that industrial leadership creates a kind of human and physical capital infrastructure, usually reflected in vigorous R&D. This intellectual property is partly external to the individual competitors in the leader economies and acts as a catalyst for new processes and products. There is a virtuous circle at work so long as resources can move freely from yesterday's staples to tomorrow's winners. When structural adaptation is slow, international competition can remove a historical leadership position from an economy whose high relative wages then prevent it from competing later in the cycle with newly-industrializing fast-followers. The UK is an economy, open to fierce international competition, which has lost its place in this way and is rapidly de-industrializing.

The typical market in international trade nowadays is one in which there is a small number of firms each with a strong home base but competing in each other's markets and in third markets. Price cuts are typically quickly matched, so oligopolists compete (and try to secure their segment of the market) by differentiating their products. Cars, clothes and computers are therefore both imported and exported as international competition drives individual producers to specialize.

Much of the most recent work on international competition examines how trade develops under conditions of imperfect competition. Intra-industry trade flows are affected by increasing returns to scale in individual product lines, by product differentiation, by competition in *non-price* terms (design, reliability, durability, variety, packaging, servicing, distribution, delivery speed, advertising) and by the scope for vertically integrated corporations to choose the separate locations of the different stages of production of the final product.

With intra-industry trade specialization, the threat of re-entry into each other's specialisms, as well as the possibilities of new entry, limit monopoly power. Other things being equal, entry is easier where world demand is growing fast and when the background technology is changing fast. Trade under conditions of imperfect competition tends to lead to increasing 'narrow market' specialization but decreasing 'wide market' monopoly power, together with lower costs and equal or greater breadth of consumer choice.

Two final international competition issues are important: the occasionally perverse effects of increased price competitiveness through currency depreciation; and the effect upon trade flows of the competitive options open to MNCs.

Suppose, through successive devaluations, one country's relative unit labour costs fall. (Relative unit labour costs are a better measure of price competitiveness than relative export prices because the latter catch only actual transactions, not lost orders.) The benefit to its trade in manufactures may be short-lived. Orders will come in at

the most price-sensitive end of each of its industrial markets. Since this is usually the less sophisticated end, the nation's industries tend to become less competitive in *non-price* terms: to be impelled to specialize in those product versions whose demand grows least fast as incomes grow. In the longer run, scale economies (and overall market share) may be lost. Paradoxically, devaluation may have its best chance of increasing trade competitiveness when *domestic* competition is limited, so that home currency prices can be raised in price-inelastic markets, and the higher margins be used to overcome *non-price* disadvantages in world markets.

Large firms can move technology, capital and management skills, as well as goods, and change domestic factor supplies by training. International competition drives firms to look for the lowest cost way of creating value-added. It is often cheaper for a MNC to transfer key inputs than to ship final products. Exporting, licensing, local assembly or packaging, joint ventures with local producers, and complete local production and marketing are alternative modes of international competition. Each mode has private benefits and costs. The choice between them is affected by the expected reactions of competitors and by the policies of the consumer governments. Protection, and its threat, and dramatic improvements in communications in recent years have led to a large increase in international competition between local subsidiaries of MNCs and an increase in the proportion of trade in intermediate goods within vertically integrated firms. The overall volume of trade is nowadays a poor measure of the importance or strength of international competition.

See Also

► International Trade

References

Caves, R.E. 1971. International corporations: The industrial economics of foreign investment. *Economica* 38: 1–27.

Caves, R.E., and R.W. Jones. 1981. *World trade and payments: An introduction*, 3rd ed. Boston: Little Brown.

Cornwall, J. 1977. *Modern capitalism, its growth and transformation*. London: Martin Robertson. ch. 10.

Grubel, H., and P. Lloyd. 1975. *Intra-industry trade: The theory and measurement of international trade in differentiated products*. New York: Wiley.

Houthakker, H.S., and S.P. Magee. 1969. Income and price elasticities in world trade. *The Review of Economics and Statistics* 51(2): 111–125.

Kanamori, H. 1968. Economic growth and exports. In *Economic growth: The Japanese experience since the Meiji Era*, ed. L. Klein and K. Onkawa. Homewood: Richard Irwin.

Leontief, W.W. 1934. Domestic production and foreign trade. *Economica Internazionale* 7: 3–32. Reprinted in American Economic Association, *Readings in international economics*, Homewood: Richard Irwin, 1968.

Linder, S. 1961. *An essay on trade and transformation*. Stockholm: Almqvist & Wiksell.

Schumpeter, J.A. 1942. *Capitalism, socialism and democracy*. New York: Harper.

Stout, D.K. 1977. *International price competitiveness, non-price factors and export performance*. London: National Economic Development Office.

Thirlwall, A.P. 1980. *Balance of payments theory and the UK experience*. London: Macmillan.

Venables, A.J. 1985. *International trade, trade and industrial policy and imperfect competition: A survey*. Centre for Economic Policy Research Discussion Paper No. 74, Oct.

Vernon, R. (ed.) 1970. *The technology factor in international trade*. National Bureau of Economic Research Conference Series No. 22. New York.

Wells, L. 1972. International trade: The product life cycle approach. In *The product life cycle and international trade*, ed. L. Wells. Cambridge, MA: Harvard University Press.

Competition Policy

Alan Hughes

The content and direction of competition policy is inevitably conditioned by the domestic structure and international relations of the economy in which the policy is applied. Since my discussion, although theoretical, is essentially concrete, I will direct my analysis to the case of competition policy in the UK. However, the argument may readily be generalized to other economies.

Economic Efficiency

In the postwar period there have been considerable changes in the industrial structure and performance of the UK economy. In general, these changes have been associated with a decline in international competitiveness across broad areas of manufacturing industry, and a more concentrated structure of output in the hands of fewer, major, domestic producers in many manufacturing and non-manufacturing sectors. The increased exposure of the UK to international competition following entry to the EEC, and the more general progressive liberalization of international trade in the world economy, have been associated, in this country, with the development of substantial structural unemployment, surplus capacity, and an inability to match, by exports, the increased penetration of domestic markets by foreign producers.

The policy response to these changes has involved a constant interplay between macroeconomic demand management on the one hand, and interventionist, cooperative and competitive supply side strategies on the other. As one industrial policy weapon, on the supply side, the regulation of competition has evolved gradually in the postwar years in response to the accumulation of evidence about the effects of particular types of market behaviour upon the amount, quality and price of output supplied, and upon the responsiveness of supply to changing domestic and international demand patterns. This evolution has occurred against a background of other supply side industrial policies, in particular planning and cooperative initiatives, and policies towards price and profit margin controls. Any discussion of the appropriate current, and future, stance of competition policy must therefore be carried out in the context of specific assumptions about the objectives and form of industrial policy; and about the form of overall economic strategy, in particular about policies towards the balance of payments and international trade.

In an open economy we may define the role of industrial policy to be the maintenance of efficient production of output. By efficiency in this sense we mean that the economy must be able to meet the demands for goods and services of consumers at home, as well as sell enough of its products abroad to pay for the nation's import requirements subject to the constraint that this objective is fulfilled at socially acceptable levels of output, employment and the exchange rate (Singh 1977).

It will be assumed that the role of competition policy within this framework must be to regulate economic behaviour to assist in the achievement of the efficiency objective. In more operational terms we shall interpret this to mean that competitive behaviour should be so regulated as to ensure the production, currently and in the future, of internationally tradeable output of a sufficiently high quality and at a sufficiently low cost and price to compete effectively with international suppliers at home and abroad, along with least-cost production of those non-tradeable elements of domestic consumption. This means that the competitive process must so far as is possible allocate sufficient inputs, including investment, to those uses necessary to achieve those objectives, and ensure operation at lowest possible cost. To the extent that certain forms of competitive behaviour are not compatible with the achievement of the appropriate levels and allocations of investment, employment and output, and hinder the achievement of the overall efficiency objective, then competition policy must be sufficiently flexible to accommodate non-competitive behaviour shown to be necessary for this.

I am therefore abstracting throughout from arguments in favour of a competition policy based on the notion of competitive behaviour as a 'good thing' *in itself*, irrespective of its implications in terms of an overall economic efficiency objective. Such arguments may be considered by some to be of overriding importance. It must, however, be recognized that they are based on political, philosophical or moral grounds, and are not in themselves susceptible to positive economic analysis.

Competitive and Competition Policy

In order to distinguish competitive and non-competitive behaviour, and to consider the

link between market behaviour and some notion of economic efficiency we obviously need to define a concept of competition. The best-developed notion of competition in formal economic analysis is that inherent in the state of affairs known as perfect competition. It is easy to show that, under certain very restrictive conditions, perfect competition will lead to economic efficiency in the traditional Pareto sense in which resources are so allocated that no one can be made better off without someone else being made worse off. Unfortunately, both the notion of perfect competition and the related Pareto efficiency criterion are too restrictive to serve as useful foundations for the analysis of competition policy. The state of affairs embodied in perfect competition assumes away nearly all those aspects of business behaviour with which competition policy is concerned (such as rivalry in terms of price setting, innovation in products and processes, and advertising) and ignores the static welfare gains to be had by improving the internal organization of enterprises, in favour of an analysis concentrating exclusively on the gains to be had by reallocating resources between perfectly internally efficient producers. Moreover, even within its own restricted ambit, it yields few helpful decision rules for an imperfect world to adopt since it is unfortunately the case that if the conditions necessary for perfect competition are unavoidably absent in any single market then nothing can be said in general about the correct behaviour to be followed in the rest of the economy. In particular, it does not follow, for example, that the perfectly competitive partial equilibrium solution of setting prices equal to marginal cost, so that normal profits are earned in all sectors but the affected one, is the next best solution to setting prices equal to marginal cost everywhere. This 'problem of the second best' means that appropriate pricing rules can only be derived after a piecemeal approach to individual cases in which the input–output links with the distorted sectors are examined in order to gauge in which direction, and to what extent, price should diverge from marginal cost. The Pareto efficiency criterion is, of course, limited because of its inability to deal with those economic changes which involve alterations in the distribution of income. It is inevitable that the achievement of the kind of efficiency we have described above as the objective of UK economic policy will involve changes in the distribution of income. It not least between different industrial nations, and between sectors of the domestic economy. For all these reasons we have preferred to adopt another approach to competition to use in relation to our efficiency criterion set out earlier.

This approach analyses competition, not as a state of affairs, but as a dynamic process linking structural change with market behaviour. Competition is taken to mean that range of activities aimed at meeting the objectives of one producer at the expense of others, and is thus defined in the business sense of the word, as a process involving rivalry between producers. Competitive rivalry takes both the form of contests within existing markets, and the form of potential entry into new areas when prospective returns appear relatively attractive. It includes rivalry in terms of price, but also in terms of altered or improved techniques of production or products, and in terms of the provision of information to consumers about products. All these forms of rivalry have consequences for the level and rate of growth of the technical efficiency of production and standards of consumption, for the allocation of resources between industries, and for the evolution of the structures of markets themselves.

There are a number of theories from which we may choose, which involve notions of competition as a dynamic process. The best known are those of Marx (1867–94), Schumpeter (1942), Downie (1958) and Clark (1961). This entry is based on the characterization of the competitive process by Downie. This is because he provides in a relatively simple manner, an analysis which is both more precise for our purposes than Clark, and avoids the complications imposed by the richer, more cosmic, implications of the Marxian and Schumpeterian analyses of capitalism. Moreover, the Downie model includes the main beneficial effects usually claimed to follow from competitive rivalry, in terms of cost efficiency, resource allocation, and technical progress, whilst emphasizing the possible structural changes flowing from competition which may change its

nature and effects. We can then, following Downie, sketch an outline of a simple competitive process in a given market and consider its efficiency implications.

In any market we may expect that firms with the lowest cost structure (inherited from the past). will for any given market price have the highest profits to finance expansion. In their own pricing policy, firms are expected to set a price which promises to attract customers and provide sufficient retained profits to finance the capacity necessary to serve them in a balanced manner. In this way, within markets, relatively low costs will be associated with relatively fast growth, and a competitive transfer of market shares from the less to the more efficient will occur. Moreover, this transfer mechanism may have a feedback effect on efficiency in the sense that the previously less efficient will be threatened with an ever-diminishing share of the market unless they can improve their cost position, for instance by introducing more recent innovations in production technique.

The transfer and innovation mechanisms within markets thus have beneficial effects, both upon allocative efficiency, by transferring output and resources to the currently most efficient producers, and upon technical progressiveness, by encouraging the introduction of the least-cost techniques available. The outcome of this competitive process, however, in terms of the changing allocation of output between different firms over time, and the impact of this upon competitive behaviour itself, will depend ultimately on the answer to a number of empirical questions about the relationship between past and future efficiency, and company size and performance. If past success is repeated, and resulting gains in relative size offer efficiency advantages in themselves, through economies of scale or enhanced innovative ability, then particular markets may come to be dominated by ever fewer, ever larger, firms as a result of the transfer mechanism, whilst the innovative mechanism, acting as a spur for past losers to improve performance, may not be powerful enough to offset this tendency. Past failure may raise the desire, but inhibit the ability, to recover. Innovation in process or product may be expensive, and low profits, or relatively small scale, may limit the power of losers to respond actively enough to prevent the transfer mechanism from leading to the concentration of ever more output in fewer hands. Thus, in the absence of revival the end product of the competitive process may be domination of the market by the single producer whose efficiency in the past has outstripped all rivals.

However, before this stage is reached, the private gains from collusion between remaining producers may become apparent, since it can offer both lower costs of competition and a higher market price. The costs of adding to the stock of goodwill accumulated by past advertising can be reduced, and the necessity to indulge in competitive process and product innovation is modified. Moreover, collusion can offer the leading firms higher margins by concerted pricing policies, at least in the short run, and possibly in the long run too, if the scales of production and advertising created by the competitive process itself form effective barriers to entry. Thus the competitive process seems to end inevitably in dominant-firm monopolistic or oligopolistic structures, and in the suppression of the process itself. This reduces the pressure to increase the level of efficiency through time and, in the case of collusion, to reduce the dispersion of efficiency. Market price is higher, and to some extent output is lower than it might otherwise be. High-cost plants and firms can remain in production at the higher market price and the incentive to improve products and production methods falls. Thus, both levels and rates of growth of efficiency in any market may be worsened by the cessation of the competitive process in structural conditions which may themselves limit the chances of substantial new entry.

In terms of this analysis the objective of competition policy would appear to be the identification and regulation of those structures, restrictions on entry and kinds of behaviour which arrest the competitive process, and produce the efficiency losses described above. This clearly poses problems of establishing links between structure and behaviour, of specifying behaviour inimical to the competitive process and of establishing the existence of markets where the competitive process has atrophied, and where it appears that efficiency

gains are possible. In the case of an open economy such as the United Kingdom the relevant markets, and standards of comparison for efficiency, are international in character, and we must recognize that in many industries the degree of concentration of purely domestic producers will be of little use in assessing the fierceness of the transfer mechanism, because of the importance of foreign competition.

However, more fundamental problems have also to be recognized. It may be that a competitive process of the kind outlined above may not be the only, or the most efficient, means of achieving the same ends, either before or after the stage is reached at which the possibilities of market dominance or collusion arise. Thus, we must consider planned or cooperative solutions to the problem of raising efficiency. Moreover, there is no guarantee that the end result of the competitive process itself, in terms of the levels and rates of change of operating efficiency of producers, and their distribution between activities, will, for any individual country, necessarily meet the overall efficiency objective outlined earlier, in terms of output, inflation, employment and international trade performance targets. There is then, for both the above reasons, the possibility of conflict, in industrial policy *methods*, between the promotion of arm's-length competitive behaviour through competition policy, and the possible reduction in such behaviour encouraged by other aspects of government policy in pursuit, for instance, of industrial strategy. The possibility of conflict is most obvious in relation to restrictive trade practices, but is also present in the creation or encouragement of domestic dominant-firm positions through government-sponsored mergers or rationalization schemes (NEDO 1978).

These conclusions run in some cases contrary to the majority of recent academic argument in this area, where increasing emphasis has been paid to the claimed welfare losses due to merger, increased monopoly power, and restrictive trade practices. However, in the context of current UK industrial performance and within the current overall economic policy framework a detailed scrutiny of the current evidence for the UK yields no foundation for a generally more aggressive approach towards large-firm dominance than the current legislation already adopts. This is not to deny that competition policy is necessary, but to assert that an *appropriate* competition policy must be designed in the light of existing economic conditions; of the evidence of the effects of various forms of market structure and behaviour; and of the overall objectives of economic policy.

See Also

- ▶ Anti-trust Policy
- ▶ Market Structure

Bibliography

Clark, J.M. 1961. *Competition as a dynamic process*. Washington, DC: Brookings Institution.

Downie, J. 1958. *The competitive process*. London: Duckworth.

Marx, K. 1867–94. *Capital.* London: Lawrence & Wishart, 1970–72.

NEDO. 1978. *Competition policy.* London: HMSO/ National Economic Development Office.

Schumpeter, J.A. 1942. *Capitalism, socialism and democracy.* London: George Allen & Unwin, 1968.

Singh, A. 1977. The UK industry and the world economy: A case of de-industrialisation? *Cambridge Journal of Economics* 1(2): 113–36.

Competition, Austrian

Paul J. McNulty

Keywords

Allocative efficiency; Austrian economics; Competition, Austrian; Creative destruction; Entrepreneurship; Imperfect competition; Innovation; Market power; Mises, L.E. von; Monopolistic competition; Monopoly; Perfect competition; Schumpeter, J.A

JEL Classifications

B0

The essence of Austrian economics is its emphasis on the ongoing economic process as opposed to the equilibrium analysis of neoclassical theory. Austrian concepts of competition reflect this emphasis. Indeed, one of the central challenges by Austrians to the neoclassical model, and a common denominator of virtually all Austrian economics, is the rejection of the concept of perfect competition. In this respect, a number of economists who cannot be considered Austrian in all aspects of their work, share, nonetheless, the Austrian emphasis on actual market activities and processes – for example, Joseph Schumpeter (1942), J.M. Clark (1961), Fritz Machlup (1942) and others.

When the concept of competition entered economics at the hands of Adam Smith and his predecessors, it was not clearly defined, but it generally meant entry by firms into profitable industries (or exit from unprofitable ones) and the raising or lowering of price by existing firms according to market conditions. There was little recognition, and virtually no analysis, of entrepreneurship as it might be reflected in these and other forms of competition, but there was a recognition that business firms do in most situations have some control over market prices, with the degree of control varying inversely with the number of firms in the industry. These basic ideas, expanded and supplemented, are generally compatible with most modern Austrian analysis.

What is objectionable to Austrian economists is the neoclassical concept of perfect competition, developed during the 19th and early 20th centuries. The development began with Cournot (1838), whose concern it was to specify as rigorously as possible the *effects* of competition, after the *process* of competition had reached its limits. His conceptualization of this situation was a market structure in which the output of any one firm could be subtracted from total industry output with no discernible effect on price. Later contributions by Jevons, Edgeworth, J.B. Clark and Frank Knight led to the model of perfect competition as we know it today (Stigler 1957; McNulty 1967).

The trouble with the concept from the Austrian point of view, as Hayek has emphasized, is that it describes an equilibrium situation but says nothing about the competitive process which led to that equilibrium. Indeed, it robs the firm of all business activities which might reasonably be associated with the verb 'to compete' (von Hayek 1948). Thus, firms in the perfectly competitive model do not raise or lower prices, differentiate their products, advertise, try to change their cost structures relative to their competitors, or do any of the other things done by business firms in a dynamic economic system. This was precisely the reason why Schumpeter insisted on the irrelevance of the concept of perfect competition to an understanding of the capitalist process.

For Schumpeter, any realistic analysis of competition would require a shift in analytical focus from the question of how the economy allocates resources efficiently to that of how it creates and destroys them. The entrepreneur, a neglected figure in classical and neoclassical economics, is the central figure in the Schumpeterian analytical framework. The entrepreneur plays a disequilibrating role in the market process by interrupting the 'circular flow' of economic life, that is, the ongoing production of existing goods and services under existing technologies and methods of production and organization. He does this by innovating – that is, by introducing the new product, the new market, the new technology, the new source of raw materials and other factor inputs, the new type of industrial organization, and so on. The result is a concept of competition grounded in cost and quality advantages which Schumpeter felt is much more important than the price competition of traditional theory and is the basis of the 'creative destruction' of the capitalist economic process. It produces an internal efficiency within the business firm, the importance of which for economic welfare is far greater, Schumpeter argued, than the allocative efficiency of traditional economic theory (Schumpeter 1942).

His emphasis on the advantages of the firm's internal efficiency led Schumpeter to a greater tolerance for large-scale business organizations, even for those enjoying some degree of monopoly power, than was typical of many more traditional theorists of his time. This is a not uncommon

characteristic of Austrian economics. Hayek, for example, makes the distinction between entrenched monopoly, with its probable higher-than-necessary costs, and a monopoly based on superior efficiency which does relatively little harm since in all probability it will disappear, or be forced to adjust to market conditions, as soon as another firm becomes more efficient in providing the same or a similar good or service (von Hayek 1948). And that is precisely Schumpeter's point. The ground under even large-scale enterprise is constantly shaking as a result of the competitive threat from the new firm, the new management, or the new idea. Schumpeter's competitive analysis was less a defence of monopoly power than of certain business activities which were judged to be monopolistic only from the comparative standpoint of the model of perfect competition. He insisted that the quality of a firm's entrepreneurship was of far greater significance than its mere size.

The leading contemporary Austrian theorist of competition is Israel Kirzner (1973). Kirzner's approach draws on the analysis of market processes and the concept of 'human action' developed earlier by Ludwig von Mises. For von Mises, entrepreneurship is human action in the market which successfully directs the flow of resources toward the fulfillment of consumer wants (von Mises 1949). Kirzner's more fully developed theory of competition is based on the idea that the means – end nexus of economic life is not given but is itself subject to creative human action. This creative role Kirzner defines as entrepreneurship, and it is essentially the ability to detect new but desired human wants, as well as new resources, techniques, or other ways through which to satisfy them. Whether he discovers new wants or new means of satisfying old ones, the Kirznerian entrepreneur is the one who sees and exploits what others fail to notice – the profit opportunities inherent in any situation in which the prices of factor inputs fall short of the price of the final product.

There is a difference between Kirzner's theory of entrepreneurship and that of Schumpeter. Schumpeter's entrepreneur is a disequilibrating force in the economic system; he initiates economic change. Kirzner's entrepreneur plays an equilibrating role; the changes he brings about are responses to the mistaken decisions and missed opportunities he detects in the market. Unlike Schumpeter's entrepreneur, he is not so much the creator of his own opportunities as a responder to the hitherto unnoticed opportunities that already exist in the market. Thus, in the competitive market process, the Schumpeterian and Kirznerian entrepreneurs may complement each other – the one creating change, the other responding to it.

Austrian dissatisfaction with the perfectly competitive model extends to the theories of imperfect and monopolistic competition. Hayek's and Kirzner's criticisms are the same as of perfect competition, namely, that the analysis is limited to an equilibrium situation in which the underlying data are assumed to be adjusted to each other, whereas the relevant problem is the process through which adjustment occurs. Schumpeter criticized monopolistic competition for its continued acceptance of an unvarying economic structure and forms of industrial organization. Nonetheless, the incorporation into economic theory of quality competition and sales efforts, complementing the traditional and limited focus on price competition, as well as the efforts on the part of some industrial organization specialists and institutional economists to analyse and explain actual market processes, are developments that are generally within the Austrian tradition.

See Also

- ▶ Austrian Economics
- ▶ Competition
- ▶ Creative Destruction

Bibliography

Clark, J.M. 1961. *Competition as a dynamic process*. Washington, DC: Brookings Institution.

Cournot, A.A. 1838. *Recherches sur les principes mathématiques de la théorie des richesses*. Paris: Hachette.

Kirzner, I. 1973. *Competition and entrepreneurship*. Chicago: University of Chicago Press.

Machlup, F. 1942. Competition, pliopoly, and profit. Pts. I-II. *Economica,* N.S. 9, Pt. I, 1–23, Pt. II, 153–73.
McNulty, P. 1967. A note on the history of perfect competition. *Journal of Political Economy* 75: 395–399.
Schumpeter, J. 1942. *Capitalism, socialism, and democracy*. New York: Harper & Row. 1962.
Stigler, G. 1957. Perfect competition, historically contemplated. *Journal of Political Economy* 65: 1–17.
von Hayek, F.A. 1948. The meaning of competition. In *Individualism and economic order*, ed. F.A. Hayek. London: Routledge.
von Mises, L. 1949. *Human action*. New Haven: Yale University Press.

Competition, Classical

John Eatwell

Keywords

Cantillon, R.; Capital accumulation; Classical theory of value and distribution; Competition, classical; Intrinsic value; Market price; Marx, K. H.; Natural price; Neoclassical economics; Perfect competition; Perfect liberty; Petty, W.; Quesnay, F.; Rate of profit; Ricardo, D.; Smith, A.; Turgot, A. R. J.

JEL Classifications

B0

Only through the principle of competition has political economy any pretension to the character of a science. So far as rents, profits, wages, prices, are determined by competition, laws may be assigned for them. Assume competition to be their exclusive regulator, principles of broad generality and scientific precision may be laid down, according to which they will be regulated. (Mill 1848, p. 242)

In all versions of economic theory 'competition', variously defined, is a central organizing concept. Yet the relationship between different definitions of competition and differences in the theory of value have not been fully appreciated. In particular, the characteristics of 'perfect' competition (notably the conditions which ensure price-taking) are often read back, illegitimately, into classical discussions of competition.

The mechanisms which determine the economic behaviour of industrial capitalism are not self-evident. As a form of economy in which production and distribution proceed by means of a generalized process of exchange (in particular by the sale and purchase of labour), it possesses no obvious direct mechanisms of economic and social coordination. Yet, in so far as these operations constitute a system, they must be endowed with some degree of regularity, the causal foundations of which may be revealed by analysis. The first steps in economic investigation which accompanied the beginnings of industrial capitalism consisted of a variety of attempts to identify such regularities, often by means of detailed description and enumeration, as in the works of Sir William Petty, and hence to establish the dominant causes underlying the behaviour of markets. But what was required was not simply the description and classification which precedes analysis, but abstraction, the transcendence of political arithmetic (Smith 1776, p. 501).

The culmination of the search for a coherent abstract characterization of markets, and hence the foundation of modern economic analysis, is to be found in Chapter 7 of Book I of Adam Smith's *Wealth of Nations* – 'Of the Natural and Market Price of Commodities'. In this chapter Smith presented the first satisfactory formulation of the regularity inherent in price formation. The idea, partially developed earlier by Cantillon, and by Turgot in his discussion of the circulation of money, was that

> There is in every society . . . an ordinary or average rate of both wages and profits . . . When the price of any commodity is neither more nor less than what is sufficient to pay the rent of land, the wages of labour, and the profits of stock employed . . . according to their natural rates, the commodity is then sold for what may be called its natural price.

and that

> The natural price . . . is, as it were, the central price, to which the prices of all commodities are continually gravitating. Different accidents may sometimes keep them suspended a good deal above it, and sometimes force them down somewhat below

> it. But whatever may be the obstacles which hinder them from settling in this center of repose and continuance, they are continually tending towards it. (Smith 1776, p. 65)

Thus the natural price encapsulates the persistent element in economic behaviour. And that persistence derives from the ubiquitous force of competition: or, as Smith put it, the condition of 'perfect liberty' in which 'the whole of the advantages and disadvantages of the different employments of labour and stock must … be either perfectly equal or continually tending to equality' (p. 111), for the natural price is 'the price of free competition' (p. 68).

The relationship between competition and the establishment of what Petty called 'intrinsic value' had been discussed in the works of Petty, Boisguillebert, Cantillon and Harris as the outcome of rival bargaining in price formation, competition being the greater when the number of bargainers was such that none has a direct influence on price. Quesnay expressed the formation of competitive prices as being 'independent of mens' will … far from being an arbitrary value or a value which is established by agreement between the contracting parties' (in Meek 1962, p. 90), but he did not relate the *organization of production* to the formation of prices in competitive markets. Consideration of that relationship required the development of a general conception of the role of capital, and with it the notion of a general rate of profit formed by the competitive disposition of capital between alternative investments (Vaggi 1987).

A significant step in this direction was made by Turgot, who both conceived of the process of production as part of the circulation of money:

> We see … how the cultivation of land, manufactures of all kinds, and all branches of commerce depend upon a mass of capitals, or movable accumulated wealth, which, having been first advanced by the entrepreneurs in each of these different classes of work, must return to them every year with a regular profit … It is this continual advance and return of capitals which constitutes *what ought to be called the circulation of money.* (Turgot 1973, p. 148)

and saw that the structure of investments would tend to be that which yielded a uniform rate of profit:

> It is obvious that the annual products which can be derived from capitals invested in these different employments are mutually limited by one another, and that all are relative to the existing rate of interest on money. (Turgot 1973, p. 70)

However, Turgot neither related the determination of the rate of profit to production in general – he accepted the Physiocratic idea that the incomes of the industrial and commercial classes were 'paid' by agriculture – nor developed the conceptual framework which linked the formation of prices and of the rate of profit to the overall organization of the economy. These were to be Smith's achievements:

> If … the quantity brought to market should at any time fall short of the effectual demand, some of the component parts of its price must rise above their natural rate. If it is rent, the interest of all other landlords will naturally prompt them to prepare more land for the raising of this commodity; if it is wages or profit, the interest of all other labourers and dealers will soon prompt them to employ more labour and stock in preparing and bringing it to market. The quantity brought thither will soon be sufficient to supply the effectual demand. All the different parts of its price will soon sink to their natural rate, and the whole price to its natural price. (Smith 1776, p. 65)

So in a competitive market there will be a tendency for the actual prices (or 'market prices' as Smith called them) to be relatively high when the quantity brought to market is less than the effectual demand (the quantity that would be bought at the natural price) and relatively low when the quantity brought to market exceeds the effectual demand. This working of competition was known as the 'Law of Supply and Demand'. The working of competition which constitutes the 'Law' do not identify the phenomena which *determine* natural prices. The 'Law' of supply and demand should not be confused with supply and demand 'theory', that is, the neoclassical theory of price determination which was to be developed one hundred years later. Nor should Smith's discussion of the tendencies of concrete market prices be confused with supply and demand function, which are loci of equilibrium prices.

Adam Smith's conception of 'perfect liberty' consists of the mobility of labour and stock between different uses – the mobility that is

necessary for the establishment of 'an ordinary or average rate both of wages and profits' and hence for the gravitation of market prices toward natural prices. Smith identifies four reasons why market prices may deviate 'for a long time together' above natural price, creating differentials in the rate of profit, all of which involve restriction of mobility:

(a) Extra demand can be 'concealed', though 'secrets of this kind . . . can seldom be long kept';
(b) Secret technical advantages;
(c) 'A monopoly granted either to an individual or a trading company';
(d) 'Exclusive privileges of corporation, statutes of apprenticeship, and all those laws which restrain, in particular employments, the competition to a smaller number than might otherwise go into them'.

For Smith there is some similarity in the forces acting on wages and profits which derives from his conceiving of the capitalist as personally involved in the prosecution of a particular trade or business. So the rate of profit, like the rate of wages, may be differentiated between sectors by 'the agreeableness of disagreeableness of the business', even though 'the average and ordinary rates of profit in the different employments of stock should be more nearly upon a level than the pecuniary wages of the different sorts of labour' (1776, p. 124). Landlords, capitalists and workers are all active agents of mobility. In Ricardo's discussion the emphasis shifted towards the distinctive role of capital:

> It is, then, the desire, which every capitalist has, of diverting his funds from a less to a more profitable employment, that prevents the market price of commodities from continuing for any length of time either much above, or much below their natural price. (Ricardo 1817, p. 91)

Ricardo used the term 'monopoly price' to refer to commodities 'the value of which is determined by their scarcity alone', such as paintings, rare books and rare wines (1817 pp. 249–51) which have 'acquired a fanciful value', and he argued that for 'Commodities which are monopolised, either by an individual, or by a company . . . their price has no necessary connexion with their natural value' (p. 385). His analysis of value and distribution is accordingly confined to 'By far the greatest part of those goods which are the object of desire . . . such commodities only as can be increased in quantity by the exertion of human labour, and on the production of which competition operates without restraint' (p. 12).

For Marx competition is synonymous with the generalization of capitalist relations of production. Competition is thus related to the rise to dominance of the capitalist mode of production.

> While free competition has dissolved the barriers of earlier relations and modes of production, it is necessary to observe first of all that the things which were a barrier to it were the inherent limits of earlier modes of production, within which they spontaneously developed and moved. These limits became barriers only after the forces of production and the relations of intercourse had developed sufficiently to enable capital as such to emerge as the dominant principle of production. The limits which it tore down were barriers to its motion, its development and realization. It is by no means the case that it thereby suspended all limits, nor all barriers, but rather only the limits not corresponding to it . . . Free competition is the real development of capital. (Marx 1973, pp. 649–50)

And as capitalism itself develops so does competition:

> On the one hand . . . [capital] creates means by which to overcome obstacles that spring from the nature of production itself, and on the other hand, with the development of the mode of production peculiar to itself, it eliminates all the legal and extra-economic impediments to its freedom of movement in the different spheres of production. Above all it overturns all the legal or traditional barriers that would prevent it from buying this or that kind of labour-power as it sees fit, or from appropriating this or that kind of labour. (Marx 1867, p. 1013)

The concentration of capital (increasing unit size of firms) and, in particular, the centralization of capital (cohesion of existing capitals) destroys and *recreates* competition. Competition is one of the most powerful 'levers of centralization', and

> The centralization of capitals, or the process of their attraction, becomes more intense in proportion as the specifically capitalist mode of production develops along with accumulation. In its turn

> centralization becomes one of the greatest levers of its development. (Marx 1867, p. 778n)

Like Smith and Ricardo, Marx, relates the development of competition to the establishment of the general rate of profit:

> What competition, first in a single sphere, achieves is a single market value and market price derived from the individual values of commodities. And it is competition of capitals in various spheres, which first brings out the price of production equalising the rates of profit in the different spheres. The latter process requires a higher stage of capitalist production than the previous one. (Marx 1894, p. 180)

It is in his conception of the circuit of capital that Marx best portrays capitalist competition. The image is one of capital as a homogeneous mass of value (money) seeking its maximum return. Profits are created by embodying capital in the process of production, the commodity outputs of which must be realized, that is, returned to the homogeneous money form to be reinvested. Competition is thus characteristic of the capitalist mode of accumulation; mobility and restructuring are two aspects of the same phenomenon.

Marx's general conception of capital as a system corroborates Quesnay's notion of an economy operating 'independent of men's will'. This does not mean that there may not be circumstances in which individual capitals exercise some control in particular markets – indeed, such limitations may be necessary for the accumulation process to proceed in certain lines. Capital removes only those barriers which *limit* its accumulation. The market control exercised in some lines of modern industry is not necessarily a limitation but may be a prerequisite of production on an extended scale. Aggregate capital flows discipline the actions of individual capitals, and hence endow the system with the regularity manifest in the perpetual tendency, successfully contradicted and recreated, towards a general rate of profit and associated prices.

Competition not only establishes the object of analysis, natural prices and the general rate of profit, but makes meaningful analysis possible, since it allows the operations of the capitalist economy to be characterized in a manner which permits theoretical statements of general validity to be made about them.

Theory proceeds by the extraction from reality of those forces which are believed to be dominant and persistent, and the formation of those elements into a formal system, the solution of which is to determine the magnitude or state of the variables under consideration. It is obvious that the solution will not, except by a fluke, correspond to the actual magnitudes of the variables ruling at any one time, for these will be the outcome not solely of the elements grouped under the heading 'dominant and persistent', but also of the myriad of other forces excluded from the analysis as transitory, peculiar or specific (lacking general significance) which may, at any moment, exert a more or less powerful effect. Nonetheless, the practice of analysis embodies the assumption that the forces comprising the theory *are* dominant, and that the determined magnitudes will, on average, tend to be established. In any satisfactory analytical scheme these magnitudes must be centres of gravitation, capturing the essential character of the phenomena under consideration.

The importance of Smith's use of competition is now apparent. Theory cannot exist in a vacuum. Simply labelling forces dominant is not enough. These forces must operate through a process which establishes their dominance and through which the 'law-governed' nature of the system is manifest. That process is competition, which both enforces and expresses the attempt of individual capitals to maximize profits. Thus important aspects of the behaviour of a capitalist market economy may be captured at a sufficient level of generality to permit the formulation of general causal statements, that is, to permit analysis. Without this step, which constitutes the establishment of what was called above the *method* of analysis, it would have been impossible to develop any general form of economic *theory*.

The classical theory of value and distribution may be shown to provide a logically coherent explanation of the determination of the general rate of profit and hence of natural prices (prices of production) taking as data (see Sraffa 1960):

(a) The size and composition of social output;
(b) The technique in use; and
(c) The real wage.

The classical achievement is thus composed of two independent elements: (*a*) the characterization of the object of the theory of value; and (*b*) the provision of a theory for the determination of that object. Underlying the former is the concept of gravitation imposed by competition, and underlying the latter the concept of gravitation inherent in theoretical abstraction. Any alternative system must not simply provide a different theory but also achieve a similar congruence with the traditional method.

The development in the final quarter of the 19th century of what was to become known as the neoclassical theory of value and distribution was an attempt to provide an alternative to a classical theory embroiled in the logical difficulties inherent in the labour theory of value and sullied by unsavoury associations with radicalism and Marxism. But despite the dramatic change in theory that was to be heralded by the works of Jevons, Menger and Walras, the method of analysis which characterized the object the theory was to explain stayed fundamentally the same; the new theory was an alternative explanation of the same phenomena. Marshall labelled natural prices 'long-run normal prices', and declared that, as far as his discussion of value was concerned 'the present volume is chiefly concerned … with the normal relations of wages, profits, prices etc., for rather long periods' (1920, p. 315). The same continuity of method may be found in the work of Walras (1874–7, pp. 224, 380), Jevons (1871, pp. 86, 135–6), Böhm-Bawerk (1899, p. 380) and Wicksell (1934, p. 97).

Nonetheless, the structure of neoclassical theory is such that a different notion of competition is required. The classical emphasis on mobility must be supplemented by a precise definition of the relationships presumed to exist between individual agents. The fundamental concept of 'perfect' competition, for example, encompasses the idea that the influence of each individual participant in the economy is 'negligible', which in turn leads to the idea of an economy with infinitely many participants (Aumann 1964). Such formulations are entirely absent from the classical conception of competition, since the classical theory is not constructed around individual constrained utility maximization.

See Also

▶ Competition

Bibliography

Aumann, R.J. 1964. Markets with a continuum of traders. *Econometrica* 32: 39–50.
von Böhm-Bawerk, E. 1899. *Capital and interest*. Vol. 2. South Holland: Libertarian Press. 1959.
Jevons, W.S. 1871. *Theory of political economy*. Harmondsworth: Penguin. 1970.
Marshall, A. 1920. *Principles of economics*. 8th ed. London: Macmillan.
Marx, K. 1867. *Capital*. Vol. 1. Harmondsworth: Penguin. 1976.
Marx, K. 1894. *Capital*. Vol. 3. New York: International Publishers. 1967.
Marx, K. 1973. *Grundrisse*. Harmondsworth: Penguin.
Meek, R.L. 1956. *Studies in the labour theory of value*. London: Lawrence & Wishart.
Meek, R.L. 1962. *The economics of physiocracy*. London: Allen & Unwin.
Meek, R.L. 1973. *Introduction to turgot* (1973).
Mill, J.S. 1848. *Principles of political economy*. London: Parker.
Ricardo, D. 1817. In *Principles of political economy and taxation*, ed. P. Sraffa. Cambridge: Cambridge University Press. 1951.
Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations*. *London*: *Methuen*. 1961.
Sraffa, P. 1960. *Production of commodities by means of commodities*. Cambridge: Cambridge University Press.
Turgot, A.J.R. 1973. In *Turgot on progress, sociology and economics*, ed. R.L. Meek. Cambridge: Cambridge University Press.
Vaggi, G. 1987. *The economics of françois quesnay*. London: Macmillan.
Walras, M.E.L. 1874–7. *Elements of pure economics*. Homewood: Irwin. 1954.
Wicksell, K. 1934. *Lectures on political economy*. Vol. 1. London: Routledge & Kegan Paul.

Competition: Marxian Conceptions

Willi Semmler

In the works of the classical economists such as Smith (1776) and Ricardo (1817), competition was identified as a central concept in economic theory. Free competition was regarded as the

organizing and equilibrating force in an exchange society, bringing about natural prices as centres of gravity for market prices through capital flows from areas with low rates of returns to areas with high rates. Yet compared with the theory of perfect competition, classical free competition was defined more in terms of economic behaviour than of market structure (Stigler 1957; McNulty 1968; Eatwell 1982). Marx's concept of competition, rooted in the classical theory of free competition, also refers to the behavioural activities of the capitalist firm. Marx, however, more than the classics, cast serious doubts on the stability properties of the competitive process, and he conceptualized competition as inter-firm dynamics carried out through reorganization of the firm and technical change. In this it somewhat resembles the modern theory of oligopolistic rivalry (Friedman 1982) and Schumpeter's notion of competition as a process of 'creative destruction' (1943).

The Dynamics of Competition in Marx

The marxian concept of competition though already adumbrated in his early writings (1847, 1857/8) is systematically developed in his later work (1861/3, 1867, 1893, 1894). It is derived from his theory of the behaviour of the capitalist firm (Kuruma 1973). The driving force for economic change and growth is the goal of the capitalist firm to grow and to expand ('the self-expansion of capital'). From the inter-firm dynamics results economic evolution, accumulation and growth, but also the downfall of old firms and the centralization of capital (Marx 1867, ch. 25) by which the competition and rivalry become fiercer. Firms are not conceived as powerless economic agents adjusting passively toward parametrically given techniques, prices and quantities but as actively seeking the reorganization of production and market activities in the context of rivals' possible reactions. Firms also are not seen as price takers but rather as price-setting firms with their market shares adjusting through the reaction of the rivals or as quantity-setting firms (Marx 1894, ch. 10) with prices and profits determined through market interactions. Price differentiation even for homogeneous products is assumed to exist under disequilibrium conditions (Marx 1894, ch. 10). Monopoly firms are considered exceptional cases as 'temporary monopolies' (Marx 1894, p. 178) when the demand exceeds supply for a considerable period or as 'natural monopolies' (Marx 1894, p. 861) when there is ownership of land or natural resources (Marx 1894, pp. 178, 861).

In production activities, the reorganization of the firm and technical change are seen as the main weapons of competition (Marx 1867, pp. 623). The goal of the firm is to capture a transient surplus profit and to transform it into long-run growth potentials, leading to disequilibria and imbalances through irreversible technical change and innovation, taking place not in time-continuous form but in discrete steps. Moreover, competition through technical change results not in the existence of one optimal technique but in the coexistence of multiple techniques, and the weighted average technique is – excluding some exceptional cases such as decreasing returns to scale and rent – considered the regulating technique determining the long-run normal price (Marx 1894, ch. 10).

Contrary to those forces generating disequilibria and imbalances through inter-firm dynamics competition is also conceived as a balancing force. Capital as a homogeneous fund (money capital) seeks its maximum return by flowing between sectors ('competition between industries', Marx 1894, ch. 10). Free mobility of labour and capital, no artificial or natural barriers for its entry or exit and sufficiently widespread knowledge of fields of investment are considered preconditions for the free flow of funds. In Marx (1894, ch. 10) as in Smith (1776, ch. 7), and Ricardo (1817, ch. 4) a dynamical process is conceived in which capital funds flow into industries with high rates of return away from industries with low rates of return. Thus the relative output proportions in industries will change, creating imbalances of supply and demand. These, in turn, cause relative market prices and profit rates to change, tending to establish for the economic system long-run prices of production as centres of

gravity for market prices. Yet the stability properties of such a dynamic process were not demonstrated rigorously. The arguments were put forward intuitively by analogy with Newton's theory of the planetary system that profit rates fluctuate or oscillate within a bounded interval and actual prices gravitate around their long-run production prices. Differentials of profit rates between industries and firms were expected to exist for a shorter or longer period due to disequilibrium dynamics and due to speed and ease of adjustment varying from one industry to another (Marx 1894, p. 208). Though Marx anticipated possible institutional and structural changes, due to 'concentration and centralization of capital', he did not, however, assume that inter-firm and inter-industry competition would become less severe with the evolution of capitalism.

Post-Marxian Theory

In the post-marxian theory since Hilferding (1910) the elimination of competition and the delay and disruption of the formation of a general profit rate through monopolization became the main theoretical concern. Three causes are posited as reasons for monopolization: industrial concentration, increasing constraints for the mobility of capital (in particular due to high proportion of fixed capital in total capital outlay), and collusion (cooperative behaviour and cartels). In this view these three causes result in monopoly prices and the persistence of differential profit rates between industries and size classes of firms (Sweezy 1942). For those theories, the large firms are conceived as economic units endowed with discretionary price setting power determining their own environment (Kalecki 1938, 1943; Sweezy 1939, 1942; Baran and Sweezy 1966; Eichner 1976). Here the ideas of mark-up pricing, target rate of return pricing and entry-preventing pricing have replaced the classical and marxian theory of production prices (natural prices, prices of production).

Given this general trend in post-marxian theory there are, however, many differences among theorists regarding (i) the causes of the monopolization; (ii) the determination of the mark-ups and the rates of return; (iii) the different role of inverse demand function and quantity reactions in their theory of price setting firms; (iv) the impact of the rise of oligopolies and firms size on technical change; and (v) the impact of large oligopoly firms on the stability of the economic system (increasing stability or instability with stagnation tendencies). Yet in spite of these differences, post-marxian theory is influenced by the theory of imperfect competition arising in the 1930s, and competition is thus identified more with market structure than with rival behaviour. Moreover, the theory of mark-up pricing was built more on a partial equilibrium view and thus not well-founded in an interdependent economic system. Though in the writings of some of the post-Marxian scholars the existence of large oligopoly firms does not preclude rivalry and competition (in particular concerning technical change, see Baran and Sweezy 1966, ch. 3), post-marxian writers seemed to have considered the theory of imperfect competition a more adequate framework for their analysis of advanced capitalism.

Recent Discussions

In recent discussions there is a certain revival of the concepts of competition of the classics and Marx, in particular concerning the role of competition for (1) industrial and corporate price and profit determination; (2) technical change and innovation; and (3) the formation of a general rate of profit.

1. In new contributions attempts have been made to elaborate a theory of mark-up pricing for large corporations in the context of a dynamic theory of competition and long-run prices of production. In this context the economic behaviour of large corporations is explained more in terms of change of the production processes and the organization of the firm and less in terms of a change of market structures (Clifton 1977, 1983; Semmler 1984a) as was

attempted by post-marxian theory. According to this new view, mark-up and target rate of return pricing have their origin not in new market structures but in the rise of a new type of firm: the multi-plant and multi-product corporations and their new financial management techniques. Though there is, as the theory of mark-up pricing predicts, sufficient empirical evidence of differential profit rates among industries and size classes of firms – depending, however, also on the time-period and the measure chosen for the rate of profit – it has not been sufficiently demonstrated that these differentials stem from imperfect market structures or from a disequilibrium dynamics. In addition, the empirically observed mark-up, target rate of return, and entry-preventing pricing, originally developed by large corporations in the 1920s, can be made consistent with a concept of long period prices of production. Since, however, large corporations are no longer single product firms, it is more appropriate to apply the theory of joint production to the economic behaviour of large corporations (Semmler 1984a). On this basis fruitful attempts have been made to analyse the dynamics of competition, mark-ups, and rates of return on the basis of an interdependent system of prices and outputs.

2. The theory of technical change in marxian economics has recently been given a firmer foundation in the theory of competition (Okishio 1961; Shaikh 1978; Roemer 1979). In this discussion, the marxian statement (Marx 1867, ch. 12; 1894, ch. 15) that under competitive pressure individual firms will implement technical change and innovations and capture a transient surplus profit, but that the diffusion of techniques will entail a falling general profit rate, was debated anew with the tools of mathematical economics. The Okishio theorem seemed to invalidate this statement, since it implies that the capitalist firm in competition will always choose a cost minimizing technique that raises the individual as well as the general profit rate (Okishio 1961). This Okishio result was disputed by Shaikh (1978) and extended by Roemer (1979). The latter extended the Okishio result to a production price model including fixed capital. In the debate, however, it became clear that the Okishio result holds only under the conditions of perfect competition with perfect information about the current and future cash flow and capital cost of an innovation where rivals' reactions either do not occur or can be foreseen (Semmler 1984b). In the context of the dynamics of competition as conceived by Marx – and also in the Schumpeterian tradition – due to unforeseen rivals' reactions certainty concerning future technology and markets cannot be expected when firms choose or are forced to choose a technique through competition. Thus the theory of perfect competition does not seem to be applicable as a framework in this context. But choice of techniques with market and technological uncertainties due to unforeseen rivals' reactions is by its nature difficult to model appropriately and thus more precise results are not yet available.
3. In post-marxian theories the competitive formation of a general profit rate was either taken for granted or completely disputed (as in the tradition since Hilferding). Recently, however, it became clear that if it cannot be established theoretically how profit rate differentials are dynamically equalized through the forces of competition then the concept of prices of production would become empirically irrelevant. In order to solve this problem, many scholars have begun to formalize Marx's conceptualization of competition by means of dynamical systems with price and quantity changes over time. Nikaido (1983) presented results on the dynamic equalization of profit rates and the stability properties of prices of production showing that in general they are not even locally stable. In subsequent discussion, however, it was shown by Duménil and Lévy (1984), Steedman (1984), and Flaschel and Semmler (1987) that better results may be obtained if the dynamics of competition are formalized as indicated above. For an *n*-sector

model, a dynamics which includes changes not only in prices but also in production levels can be formulated as follows:

$$\dot{x}_i = d_i\left[r(p,x)_i - \overline{r}(p,x)\right] \tag{1}$$

and

$$\dot{p}_i = k_i\left[\overline{D}(p,x)_i = S(p,x)_i\right] \tag{2}$$

where $\overline{D}(p,x)$ is the average or expected demand for an industry i for a growing economic system, $S(p,x)_i$ the industry's supply, $r(p, x)_i$ the industry's profit rate, $\overline{r}(p,x)$ the average profit rate and $\dot{p}_x$, $\dot{x}_i$ the time rate of change of prices and outputs. This dynamical process of competition refers to capital flows across industries according to differential profit rates (and changes in respective production levels) as well as to changes in prices due to imbalances in supply and demand. The results obtained in recently published articles on this process range from the demonstration of complete instability of the dynamic equalization of profit rates (Nikaido 1983) to the demonstration that prices of production are at least locally stable (Duménil and Lévy 1984). It can also be demonstrated (by utilizing a proper Lyapunov function) that prices, outputs, and profit rates are fluctuating or oscillating within boundaries (Flaschel and Semmler 1987). Most of these attempts, however, refer only to a circulating capital model when the inter-industry competitive process is analysed, and the demonstrated results depend on the type of formalization, the reaction coefficients as well as on additional stabilizing forces (such as substitution in capitalist consumption, rate of change of inventories or rate of change of profit rate differentials). Models of inter-industry competition including fixed capital, returns to scale, or multiple techniques are still rare.

See Also

- ▶ Monopoly Capitalism
- ▶ Surplus Approach to Value and Distribution

References

Baran, P., and P. Sweezy. 1966. *Monopoly capital*. New York: Monthly Review Press.

Clifton, J.A. 1977. Competition and the evolution of the capitalist mode of production. *Cambridge Journal of Economics* 1(2): 137–151.

Clifton, J.A. 1983. Administered prices in the context of capitalist development. *Contributions to Political Economy* 2: 23–38.

Duménil, G., and Lévy, D. 1984. *The dynamics of competition: A restoration of the classical analysis*. CEPREMAP, No. 8416. Paris.

Eatwell, J. 1982. Competition. In *Classical and Marxian political economy: Essays in memory of R. Meek*, ed. I. Bradley and M. Howard. London: Macmillan.

Eichner, A.S. 1976. *The Megacorp and oligopoly: Microfoundations of macrodynamics*. Cambridge: Cambridge University Press.

Flaschel, P., and Semmler, W. 1987. *Classical and neoclassical competitive adjustment processes*. The Manchester School of Economic and Social Studies.

Friedman, J. 1982. Oligopoly theory. In *Handbook of mathematical economics*, vol. II, ed. K.J. Arrow and M.D. Intriligator. North-Holland: Amsterdam.

Hilferding, R. 1910. *Das Finanzkapital*. Trans. as Finance Capital. London: Routledge & Kegan Paul, 1981.

Kalecki, M. 1938. *Distribution of national income*. Reprinted in M. Kalecki, Selected essays on the dynamics of the capitalist economy. Cambridge: Cambridge University Press, 1971.

Kalecki, M. 1943. *Cost and prices*. Reprinted in M. Kalecki, Selected essays (1971).

Kuruma, S. 1973. *Marx Lexikon zur Politischen Oekonomie*, Konkurrenz, vol. 1. Berlin: Oberbaum.

Marx, K. 1847. *Lohnarbeit und Kapital*. Trans. as Wage labor and capital. New York: International Publishers, 1933.

Marx, K. 1857/8. *Grundrisse der Kritik der Politischen Ökonomie*. Trans. as Grundrisse. New York: Random House, 1973.

Marx, K. 1861/3. *Theorien über den Mehrwert*. Trans. as Theories of surplus value, 3 vols. Moscow: Progress Publishers, 1963.

Marx, K. 1867, 1893, 1894. *Das Kapital, Kritik der Politischen Ökonomie*. 3 vols. Trans. as Capital, a critique of political economy. New York: International Publishers, 1967.

McNulty, P.J. 1968. Economic theory and the meaning of competition. *Quarterly Journal of Economics* 82: 639–656.

Nikaido, H. 1983. Marx on competition. *Zeitschrift für Nationalökonomie* 43(4): 337–362.

Okishio, N. 1961. Technical changes and the rate of profit. *Kobe University Economic Review* 7: 85–99.

Ricardo, D. 1817. Principles of political economy and taxation. In *Works and correspondence*, ed. P. Sraffa, vol. 1. Cambridge: Cambridge University Press, 1951.

Roemer, J.E. 1979. Continuing controversy on the falling rate of profit: Fixed capital and other issues. *Cambridge Journal of Economics* 3(4): 379–398.
Schumpeter, J. 1943. *Capitalism, socialism and democracy*. London: George Allen & Unwin.
Semmler, W. 1984a. *Competition, monopoly and differential profit rates*. New York: Columbia University Press.
Semmler, W. 1984b. Marx and Schumpeter on competition, transient surplus profit and technical change. *Economie Appliquee* 37(3–4): 419–455.
Shaikh, A. 1978. Political economy and capitalism: Notes on Dobb's theory or crisis. *Cambridge Journal of Economics* 2(2): 233–251.
Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations*, ed. E. Cannan. London: Methuen, 1961.
Steedman, I. 1984. Natural prices, differential profit rates, and the classical competitive process. *The Manchester School of Economic and Social Studies* 52(2): 123–140.
Stigler, G.J. 1957. Perfect competition, historically contemplated. *Journal of Political Economy* 65(1): 1–17.
Sweezy, P.M. 1939. Demand under conditions of oligopoly. *Journal of Political Economy* 47: 568–573.
Sweezy, P. M. 1942. *The theory of capitalist development*. New York: Monthly Review Press, 1968.

Competitive Market Processes

J. A. Clifton

1. Do fully competitive price signals from intense rivalry in the market justify the moral sentiment of laissez-faire? On grounds of distributive justice among risk-takers, the answer has generally been 'yes' throughout the history of economic analysis.

   In considerations of optimum economic efficiency, however, the answer seems to have become more difficult over the course of development. The cheapening of commodities witnessed by the classical economists is the most virtuous example of efficient competitive market processes in which the distribution of returns tends to be equalized. Against the theoretical standard of perfect competition, non-price forms of competition came to be viewed in the 1930s as second-class virtues – imperfectly or monopolistically competitive practices.

   Questions of intervention on grounds of allocative inefficiency have continued to hinge on the existence of classical monopoly profits. By the 1970s, the weight of empirical evidence and the acknowledged fact of intensified global competition served to eliminate the credibility of the market concentration doctrine derived from perfect and imperfect competition (see Demsetz 1973, 1982).

   At a more fundamental level, Joan Robinson was persuaded to abandon imperfect competition in favour of trying to more fully develop a classical line of analysis only partially worked out by the classical economists and largely ignored ever since (see Clifton 1977).

   Yet, after a decade of deregulation and the strongest sentimentality to let the free market reign, evidence of static and dynamic inefficiencies in industry is accumulating (*Business* Week 1986). Has the sheer intensity of competition in the rate of economic change and the pace of economic life become so severe as to hinder economic efficiency even under the strongest possible tendencies to equalized returns in the market? Competition, however complex and full of discontinuities, is still evident as a systematic and general force in the empirically observed fact that accounting rates of return across firms and industries tend toward uniformity over time.

   This dynamic tendency is stronger among larger than smaller firms and is stronger today than a century ago (see Singh and Whittington 1968 chapter 6; Brozen 1970, 1971, 1982, pp. 239–40). But it is not explained by the neoclassical theory of perfect competition, which requires atomism of independent agents under static premises of maximization. It is not explained by imperfect or monopolistic competition for stable positions of some degree of monopoly power are less and less in evidence all the time. Yet this dynamic tendency is not associated with any optimum or unique state of industrial efficiency, as under perfect competition. Finally, the intensity of competitive

rivalry that leads to this tendency cannot be measured by neoclassical standards – the number of firms in a market. It exists primarily under market conditions of concentrated oligopoly.

It seems pointless to try to reconstitute the general theory of competitive value by still more a priori game theorizing which only adds to the false perception of indeterminacy and lack of systematic generality in 'price behaviour' under contemporary market conditions. A recent alternative has been to apply game theory to perfect competition (see Mas-Colell 1980). What used to be a static state of affairs distinguished by the absence of any and all rivalry is now a non-cooperative equilibrium, independent of the number of agents, that may entail dynamic strategies of M periods contingent on past histories.

This re-introduces the long forgotten classical principle that interdependent, dynamic rivalries are what lead to the tendency toward uniformity in returns across the price system. A possible virtue of the approach is that not all games need have positive sum outcomes, so the question of competitive rivalry and economic efficiency is left open, not closed as in the pure neoclassical doctrine of perfect competition.

With all the intellectual baggage imposed by perfect and imperfect competition, however, is it not preferable to start fresh by examining and explaining in classical price–theoretic terms that systematic empirical tendency toward uniformity in returns? The first point is that the institutional conditions for free capital mobility in the industrial context of fixed capital have developed gradually and progressively over the course of economic development during the past two hundred years. They are to be found in the first instance not in the atomistic enterprise but in the evolution of the organizational structure and competitive strategies of today's representative firm, the industrially and geographically diversified, publicly held corporation. Top management in industry has increasingly assumed the role once reserved for bankers in day-to-day affairs, moving capital from areas of lower to areas of higher returns.

When finance is committed to industry as fixed capital, it is at once immobilized for its economic life. It does not have the character of putty which enables it to be moulded for any use promising today's highest return in the market. The greatest barriers to capital mobility existed for the single factory enterprise which typified organizational structures in the United States in the 1840s. Railroad firms created the first degree of capital mobility directly within the enterprise by pioneering the coordinated, multi-unit organization. From the 1890s on, a degree of capital mobility across industries was added by integrated manufacturing companies and by mass retailers. Truly diversified industrial corporations began appearing in the 1920s and by the 1930s, mass retailers were national in scope (see Chandler 1977, for a definitive history).

Beyond these structural elements in the development of capital mobility in the firm, the number of competitive strategies available to it from economies of large-scale organization and the intensity of the search for competitive advantage available from large budgets and staffs have also increased. Product innovations from a permanent R&D staff, advertising campaigns, takeovers and divestitures, together with price and credit competition give the firm added flexibility in responding to changes in market conditions and in initiating them.

Free capital mobility is not synonymous with the ability of atomistic firms or individual agents to move freely throughout the economy, whatever utopian analogies with a system of perfect liberty and individual freedom that may conjure up. What matters is the freedom of capital, however organized, so to move. As theoretical constructs, perfect and imperfect competition left a vision of capitalist development that is at complete odds with the actual historical development of conditions of free capital mobility. In this view, which is also espoused by many non-neoclassical economists, barriers to free capital mobility have

grown with the evolution of large corporations, and the system has become less competitive, not more competitive.

Even beyond considerations of corporate organization and strategy, free capital mobility is nowhere more fully developed in history than in the institutions of today's capital markets. Ever more integrated on a world scale, ever more innovative in the range of 'products' and services offered, the large firms which dominate these markets have such powerful and all-encompassing information networks as to approximate the economic assumption of perfect information in the short run, if not rational expectations in the long run.

The acceptance of market processes in ever more spheres of human existence beyond basic needs is a third sense in which free capital mobility is more highly developed today. Scale economies in automobile production are not barriers to entry into new fields of endeavour like the child care industry. Finally, with the growth of labour-intensive services as a proportion of the economy, more businesses take on the characteristics of merchant capital once again in history since even learned human capital is more malleable than fixed stock.

2. Beyond considerations of free capital mobility in explaining the uniformity in returns are other key issues that fall outside the scope of perfect and imperfect competition, whether or not amalgamated with game theory.

If today's oligopolistic firms are the slaves of the market as never before in history, in what sense are they 'pure price takers'? Such corporations are entirely unable to dictate their ex post rate of return in the market, whatever their ex ante pricing behaviour. It is with the ex post rate of return that the theory of competitive price is concerned, and that will be determined by many forms and intensities of competitive behaviour in the market, of which a suggested mark-up ex ante is only one. Partial equilibrium mark-up theories have never comprehended the difference between ex ante and ex post and err in believing ex ante pricing discretion implies some degree of ex-post monopoly power.

The very interdependence in decision-making between oligopolistic firms is what causes that ebb and flow of business and profits across firms, industries and markets so as to render the ex post rate of return fully competitive and beyond the control of the individual firm. Unfortunately, game theory was used for decades to deny the generality of contemporary competitive behaviour rather than to explain its most systematic feature in the convergence of accounting rates of return over the long run.

A virtue of the neoclassical theory of perfect competition was to provide a readily quantifiable means of measuring the intensity of competition – by the number of firms in a market. In consideration of non-price forms of competition, this precision in economic theory became lost, appearing in lieu of theory as an industrial organization 'paradigm' of market structure *and* conduct *and* performance. Can quantitative precision be resurrected in a general theory of competitive value for the modern age?

Observation tells us that the intensity of rivalry in contemporary markets can be measured by the *frequency* and *voracity* of changes in market conditions – the sum total of strategic moves and countermoves made by firms in that market per unit of time. The common denominator among all types of competition is to what measured degree does the action move business and profits from one sphere to another or one firm to another.

There is a clear analogy to perfect competition that can be made here. Were oligopolists in a market limited to the type of action an atomistic firm entering that market could take, pure price taking behaviour would emerge as the frequency of such strategic moves and countermoves increased without limit. Price for a homogeneous good would be bid down to its normal competitive minimum not by unlimited entry by one small firm after another, but by an unlimited number of atomistic-like strategic moves by the competing oligopolists.

Game theory to date appears to have overlooked the primacy of numbers of actions in

the marketplace over numbers of actors in resurrecting the general theory of competitive value, on measuring the intensity of competition as the frequency of strategic moves and countermoves in the first instance.

Of course, large firms are not restricted to atomistic competition. Cut-rate 'two percent' financing by General Motors Corporation in August 1986 was a competitive move that had the potential to draw a great amount of business and profits away from other firms. For that reason, this voracious move was imitated quickly by Ford and Chrysler.

Competition in the personal computer market has been intense not only because of voracious price breaks from time to time, but because the frequency of changes in market conditions has been enormous from real and cosmetic innovations in hardware and software. The frequency of competition among the commercial television networks in changing the time slots of programmes has at times approached the irrational from the consumer's standpoint.

3. When market processes are intensely competitive in the frequency, voracity or complexity of strategic moves and countermoves applied, what will be the nature of decisionmaking by the individual firm? Does active rivalry in the market necessarily mean 'maximizing behaviour', optimally efficient performance from decision-making at the margin?

One strong clue to the answer is the rejection of the marginal method *and* the assumption of constant returns to scale in recent classical general equilibrium models of competitive price determination (see Sraffa 1960). If maximizing behaviour underlies the classical approach, it certainly is not of pure neoclassical vintage, for decision-making at the margin requires marginal units which, according to Sraffa, are 'nowhere to be found' in the pure classical theory of competitive value.

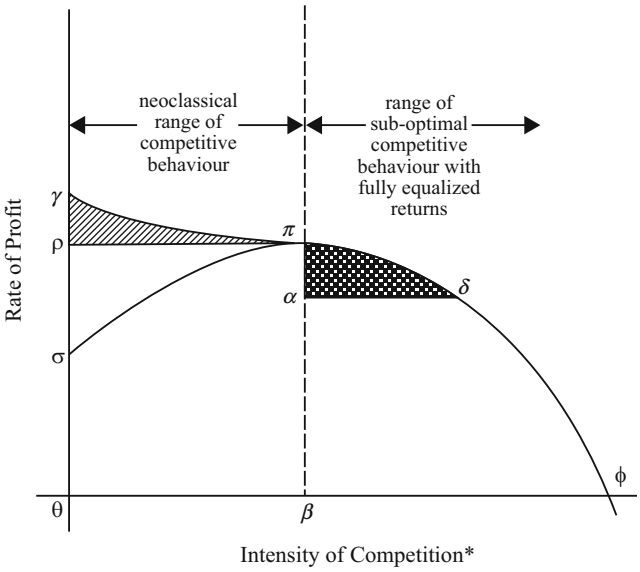
Nor is any notion of maximization or optimal efficiency to be found in the statement of technology or 'production function' of the pure classical system. The technology is not specified by input–output coefficients, which imply minimum input per unit of output. Only viability conditions for each industry at a given scale of output are listed. Viability is not the same thing as optimum efficiency in the use of a technique of production, whether under conditions of simple reproduction or the production of a surplus.

The entirely unsophisticated requirements for specifying technology in the classical determination of competitive value is an advantage, because it formally leaves open the question of whether fully competitive price behaviour in ongoing market processes is always efficient.

The empirically observed tendency of accounting rates of return to converge in the long run seems more assuredly decision-making by oligopolists where the intensity of competitive behaviour is asymmetric around a normal or average rate of profit. Whether from creditor or stockholder admonition, team pride or the threat of takeover, firms whose performance is below the normal rate of return are under stronger pressure to improve profitability than those whose performance is above the norm (see Cyert and March 1956).

Further, the attributes of intensely competitive market processes cause decision-making by the firm facing such discontinuities and complexities in its external environment to be the kind of 'bounded rationality' highlighted in the administrative theories of decision-making for different reasons related to the internal characteristics of large organizations (see Simon 1945).

The paradox of how 'maximum' effort or greater and greater rivalry directed through market processes can result in sub-optimal outcomes is precisely the question the business world, especially in America, seems to be asking itself today (see President's Commission 1985). While associated with even stronger and faster movements to capture new markets or eliminate excess profitability than less intensely competitive behaviour in bygone eras, classifying it as 'maximizing behaviour' or 'satisficing' can only lead to confusion. The



Competitive Market Processes,
Fig. 1 Efficiency and Inefficiency in Competitive Resource Allocation *For neoclassical theory, this axis measures the number of competitors, where the limit, represented by the vertical dashed line, is the familiar large numbers case of atomistic or perfect competition. For classical theory, the axis measures the frequency and magnitude of changes in market conditions. There is no limit to the intensity of such competitive behaviour

former implies efficiency where no such implication is warranted a priori, while the latter implies an absence of highly energetic behaviour from constantly striving, an implication at complete odds with the facts. A more neutral term like 'competitive behaviour' seems preferable.

4. If fully competitive price signals can exist under different degrees of industrial efficiency, then the moral sentiment of laissez-faire is not so readily justified in a competitive free enterprise system. Welfare economics must focus on competition as both virtue and vice, rather than competition as virtue and monopoly as vice, as in the past fifty years. Consider Fig. 1, which relates the intensity of competition to the degree of economic efficiency. In modern economic doctrine there are three unambiguous situations: pure monopoly (point γ), perfect competition (point π) and the long run shutdown point beyond which a firm cannot cover its total costs (point θ).

In the context of a single industry, ruinous competition is rightly viewed as self-correcting by market forces. Therefore, the entire scope of economic investigation is believed to have been between θ and β. The curve $\sigma\,\pi$ expresses the sentiment that the more competition the better for efficiency as measured by the rate of return. The curve $\gamma\,\pi$ expresses the proposition that the more competition, the lower the degree of monopoly, and the stronger the tendency toward uniformity in returns around a normal rate of profit ρ. All inefficiency is due to the absence of competition in sufficient degree, and may be measured as social welfare losses like the area $\gamma\,\sigma\,\pi$.

The principle justification for laissez-faire through history has been that 'competition without limit' must always enhance the general welfare by improving static or dynamic efficiencies, as expressed in the positive slope of $\sigma\pi$. Competition in effect can never become so intense, or of a character or complexity, that it pushes a market or an economic system beyond point π in the long run. In neoclassical theory, this is expressed as an increase in the number of firms without limit tending to produce a state of perfect competition.

Yet once we admit that ruinous competition has existed in history, is there no range of sub-optimal competitive behaviour between π and φ? Competition that is sufficiently intense to bid away all excess profits, but too intense to maximize efficiency and the general welfare? Fully competitive market processes that lead to sub-optimal outcomes – zero sum or even negative sum games?

If and only if such business practices are isolated in one or a few markets will they be self-correcting by the market. If they are, or have become, systemic throughout the economy, there is no reason to believe they will be self-regulating in the market in a way which leads to movement from a position like δ to the unique point of optimum efficiency associated with equality of returns, π.

I submit that today's general competitive equilibrium in resource allocation lies at a point like δ and that the free enterprise system in an atmosphere of laissez-faire is experiencing social welfare losses of the form $\pi \propto \delta$, not of the form $\pi \rho \gamma$ from monopolistic distortions.

There is no distortion in price signals associated with contemporary social welfare losses. They exist in a climate of intensely competitive market processes where the tendency toward equality of returns is stronger, not weaker. The real issue is becoming whether all this incessant change still represents a Schumpeterian process of creative destruction or an inefficient process of 'destructive creations'.

Free capital mobility has become so highly developed in financial markets and top management behaviour in corporations that it has led to the virtual collapse of the long period in setting aspiration levels for the rate of return on real capital formation in industry. This increase in the intensity of competition is generating an ongoing bias against efficiency-enhancing forms of strategic corporate behaviour in favour of stop-gap or crisis management forms of competition such as 'asset juggling', which does not affect the quality of products or the efficiency with which they are produced, distributed and sold.

The rate of change in and complexity of market conditions to which the firm must respond strategically has accelerated, not only in product and input markets, but also in economic policy variables here and abroad. The intensity of these competitive pressures is leading to the creation of corporate cultures that are very risk averse, and to decision-making of strictly bounded rationality that, however energetic, can hardly be called 'maximizing behaviour'.

The growing inability to protect positions of differential rent or supra-normal profits for a period necessary to sustain some of the most productive forms of risk-taking entrepreneurial behaviour is caused by the very intensity of competition in contemporary market processes. The crowding out of these Schumpeterian forms of dynamically efficient market processes is a third social welfare loss that exists in today's laissez-faire atmosphere.

The capitalization of finance on pure finance rather than real asset creation has become almost an epidemic of market processes that are of dubious value to the general welfare and that, moreover, increase the cost of capital for productive uses. For example, the increase in takeover divestiture type activities is associated with the creation of a distinct market for corporate control which simply changes the distribution of ownership and/or control of existing productive assets.

5. All seem to be agreed that competition has become more intense in recent decades and especially in recent years. I continue to maintain, as well, that there has been a secular increase in the intensity and complexity of competition over the course of capitalist development and that the free enterprise system continues to develop fundamentally along the lines of ever greater capital mobility.

   But it is also my contention that over the course of capitalist development and especially evident in recent years in America, the intensity of competition has become so great as to hinder industrial efficiency. Change for the sake of change rather than for economic and social progress. Competition, that engine of prosperity that has propelled us forward for two centuries, now seems to be of a character that it is holding us back.

This suggests a very different role for economic doctrine and public policy than either laissez-faire or the regulation of monopolistically competitive practices. It implies that intervention in the market which reduces the intensity or scope of certain

fully competitive practices will not inexorably lead to protected positions of monopoly or associated inefficiencies. Intervention may in all probability enhance economic growth or improve statical resource allocation while fully maintaining that attribute of distributive justice among risk-takers, insofar as the equality of returns is concerned, that is the hallmark of capitalism and freedom.

See Also

- ▶ Competition
- ▶ Competition, Austrian
- ▶ Competition, Classical
- ▶ Competition: Marxian Conceptions

References

Brozen, Y. 1970. The antitrust task force deconcentration recommendations. *Journal of Law and Economics* 13(2): 279–292.

Brozen, Y. 1971. The persistence of 'high rates of return' in high stable concentration industries. *Journal of Law and Economics* 14(2): 501–512.

Brozen, Y. 1982. *Concentration, mergers and public policy*. New York: Macmillan.

Business Week. 1986. The hollow corporation. 3 Mar.

Chandler Jr., A.D. 1977. *The visible hand*. Cambridge, MA: Harvard University Press.

Clifton, J.A. 1977. Competition and the evolution of the capitalist mode of production. *Cambridge Journal of Economics* 1(2): 137–151.

Cyert, R., and J. March. 1956. Organizational factors and the theory of oligopoly. *Quarterly Journal of Economics* 70: 44–64.

Demsetz, H. 1973. *The market concentration doctrine*. American Enterprise Institute for Public Policy Research, Hoover Institution on War, Revolution and Peace.

Demsetz, H. 1982. *Economic, legal, and political dimensions of competition*. New York: North-Holland.

Mas-Colell, A. 1980. Noncooperative approaches to the theory of perfect competition: Presentation. *Journal of Economic Theory* 22(2): 121–135.

President's Commission on Industrial Competitiveness. 1985. *Global competition: The new reality*, vol. 1. Washington, DC: U.S. Government Printing Office.

Simon, H. 1945. *Administrative behavior*. New York: Macmillan and Free Press.

Singh, A., and G. Whittington. 1968. *Growth, profitability and valuation*. Cambridge: Cambridge University Press.

Sraffa, P. 1960. *Production of commodities by means of commodities*. Cambridge: Cambridge University Press.

Computation of General Equilibria

Herbert E. Scarf

Abstract

The Walrasian model of economic equilibrium is a generalization to the entire economy of the basic notion that prices move to levels that equilibrate supply and demand. Although the model avoids some factors of economic significance, it is extremely useful in helping us evaluate the effects of changes in economic policy or the economic environment. A moderately realistic model designed to illustrate a significant economic issue typically involves a large system of highly nonlinear equations and inequalities. Existence of a solution is demonstrated by non-constructive fixed point theorems. The explicit numerical solution of such a model requires sophisticated computational techniques.

Keywords

Arrow–Debreu model; Barone, E.; Brouwer's fixed-point theorem; Cobb–Douglas function; Computation of general equilibria; General equilibrium; Harberger, A.; Johansen, L.; Kakutani's fixed-point theorem; Kuhn–Tucker Theorem; Lange, O.; Non-convexity; Sperner's lemma; Technical coefficients of production; Uncertainty; Walras's Law; Walrasian model

JEL Classifications

C68

The general equilibrium model, as elaborated by Walras and his successors, is one of the most comprehensive and ambitious formulations in the current body of economic theory. The basic ingredients with which the Walrasian model is constructed are remarkably spare: a specification of the asset ownership and preferences for goods and services of the consuming units in the

economy, and a description of the current state of productive knowledge possessed by each of the firms engaged in manufacturing or in the provision of services. The model then yields a complete determination of the course of prices and interest rates over time, levels of output and the choice of techniques by each firm, and the distribution of income and patterns of saving for each consumer.

The Walrasian model is essentially a generalization, to the entire economy and to all markets simultaneously, of the ancient and elementary notion that prices move to levels which equilibrate supply and demand. No intellectual construction of this scope, designed to address basic questions in a subject as complex and elusive as economics, can be described as simply true of false – in the sense in which these terms are used in mathematics or perhaps in the physical sciences. The assertions of economic theory are not susceptible to crisp and immediate experimental verification. Moreover, the Walrasian model disregards obvious aspects of human motivation which are of the greatest economic significance and which cannot be addressed in the language of our subject: economic theory is mute about our affective lives, about our opposing needs for community and individual assertion, and about the non-pecuniary determinants of entrepreneurial energy.

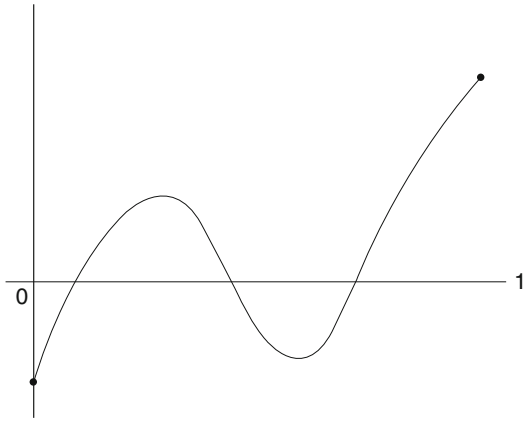
There are, in addition, aspects of economic reality which are capable of being described in the framework of the Walrasian model but which must be assumed away in order for the model to yield a determinate outcome. Uncertainty about the future is an ever-present fact of economic life, and yet the complete set of markets for contingent commodities required by the Arrow–Debreu treatment of uncertainty is not available in practice. Economies of scale in production are a central feature in the rise of the large manufacturing entities which dominate modern economic activity; their incorporation into the Walrasian model requires the introduction of non-convex production possibility sets for which the competitive equilibrium will typically fail to exist.

In spite of its many shortcomings, the Walrasian model – if used with tact and circumspection – is an important conceptual framework for evaluating the consequences of changes in economic policy or in the environment in which the economy finds itself. The effects of a major shock to the economy of the United States – such as the four-fold increase in the price of imported oil which occurred in late 1973 – can be studied by contrasting equilibrium prices, real wages and the choice of productive techniques both before and after the event in question. Generations of economists have used the Walrasian model to analyse the terms of trade, the impact of customs unions, changes in tariffs and a variety of other issues in the theory of International Trade. And much of the literature in the field of Public Finance is based on the assumption that the competitive model is an adequate description of economic reality.

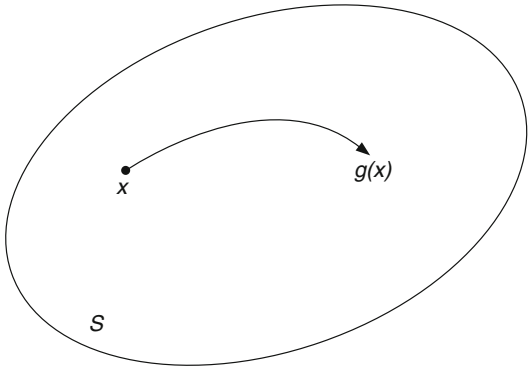
In these discussions the analysis is frequently conducted in terms of simple geometrical diagrams whose use places a severe restriction on the number of consumers, commodities and productive sectors that can be considered. This is in contrast to formal mathematical treatments of the Walrasian model, which permit an extraordinary generality in the elaboration of the model at the expense of immediate geometrical visualization. Unfortunately, however, it is only under the most severe assumptions that mathematical analysis will be capable of providing unambiguous answers concerning the direction and magnitude of the changes in significant economic variables, when the system is perturbed in a substantial fashion. In order for a comparative analysis to be carried out in a multi-sector framework it is necessary to employ computational techniques for the explicit numerical solution of the highly non-linear system of equations and inequalities which represent the general Walrasian model.

The Use of Fixed-Point Theorems in Equilibrium Analysis

One of the triumphs of mathematical reasoning in economic theory has been the demonstration of the existence of a solution for the general equilibrium model of an economy, under relatively mild assumptions on the preferences of consumers and the nature of production possibility sets (see



Computation of General Equilibria, Fig. 1



Computation of General Equilibria, Fig. 2

Debreu 1982). The arguments for the existence of equilibrium prices inevitably make use of Brouwer's fixed-point theorem, or one of its many variants, and any effective numerical procedure for the computation of equilibrium prices must therefore be capable of computing the fixed points whose existence is asserted by this mathematical statement.

Brouwer's fixed-point theorem, enunciated by the distinguished Dutch mathematician L.E.J. Brouwer in 1912, is the generalization to higher dimensions of the elementary observation that a continuous function of a single variable which has two distinct signs at the two endpoints of the unit interval, must vanish at some intermediary point. In Brouwer's Theorem the unit interval is replaced by an arbitrary closed, bounded convex set S in R^n, and the continuous function is replaced by a continuous mapping of the set S into itself: $x \rightarrow g(x)$ Brouwer's Theorem then asserts the existence of at least one point x which is mapped into itself under the mapping; that is, apoint x for which $x = g(x)$.To see how this conclusion is used in solving the existence problem let us begin by specifying, in mathematical form, the basic ingredients of the Walrasian model (Fig. 1).

The typical consumer is assumed to have a preference order for, say, the non- negative commodity bundles $x = (x_1,x_2,\ldots,x_n)$ in R^n; the preference ordering is described either by a specific utility function $u(x_1,x_2,\ldots,x_n)$ or by means of an abstract representation of preferences. The consumer will also possess, prior to production and trade, a vector of initial assets $w=(w_1,w_2,\ldots,w_n)$. When a non- negative price vector $p= (p_1,p_2,\ldots, p_n)$ is announced the consumer's income will be $I = p \cdot w$ and his demands will be obtained by maximizing preferences subject to the budget constraint $p \cdot x \leq p \cdot w$. If the preferences satisfy sufficient regularity assumptions, the consumer's demand functions $x(p)$ will be single-valued functions of p, continuous (except possibly when some of the individual prices are zero), homogeneous of degree zero and will satisfy the budget constraint $p \cdot x(p) = p \cdot w$ (Fig. 2).

The market demands are obtained by aggregating over individual demand functions and, as such, will inherit the properties described above. The market *excess* demand functions, which I shall denote by $f(p)$, arise by subtracting the supply of assets owned by all consumers from the demand functions themselves. It is these functions which are required for a complete specification of the consumer side of the economy in the general equilibrium model: they may be obtained either by the aggregation of individual demand functions – as we have just described – or they may be directly estimated from econometric data. The following properties will hold, either as a logical conclusion or by assumption:

1. $f(p)$ is homogeneous of degree zero.
2. $f(p)$ is continuous in the interior of the positive orthant.
3. $f(p)$ satisfies the Walras Law $p \cdot f(p)=0$.

The first of these properties permits us to normalize prices in any one of several ways; for example, $\Sigma pj = 1$ or $\sum p_j^2 = 1$. Given either of these normalizations, I personally do not find it offensive to extend the property of continuity to the boundary, even though there are elementary examples of utility functions, such as the Cobb–Douglas function, for which this would not be correct.

The production side of the economy requires for its description a complete specification of the current state of technical knowledge about the methods of transforming inputs into outputs – with commodities differentiated according to their location and the time of their availability. This can be done by means of production functions, an input/output table with substitution possibilities and several scarce factors rather than labour alone, or by a general activity analysis model:

$$\mathrm{A} = \begin{bmatrix} -1 & 0 & \dots & 0 & a_{1,n+1} & a_{1,k} \\ 0 & -1 & & 0 & a_{2,n+1} & a_{2,k} \\ \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & & -1 & a_{n,n+1} & a_{n,k} \end{bmatrix}$$

Each column of A describes a particular productive process, with inputs represented by non-negative entries and outputs by positive entries. The activities are assumed capable of being used simultaneously and at arbitrary non-negative levels $x = (x_1,x_2,\dots,x_k)$; the net production plan is then $y = \mathrm{A}x$ (Fig. 3).

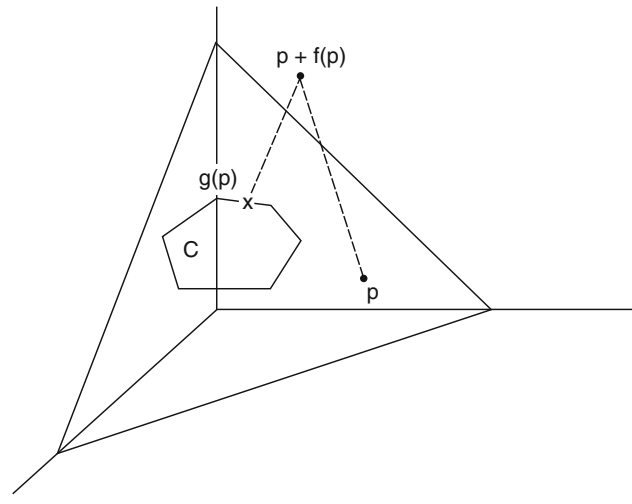
With this formulation, a competitive equilibrium is defined by a non-negative vector of prices $p = (p_1,p_2,\dots,p_n)$ and a non-negative vector of activity levels $x = (x_1,x_2,\dots x_k)$ satisfying the following conditions:

1. $f(p) = \mathrm{A}x$,
2. $p\mathrm{A} \leq 0$.

The first condition states that supply and demand are equal in all markets, and the second that there are not opportunities for positive profits when the profitability of each activity is evaluated at the equilibrium prices. Taken in conjunction with the Walras's Law, these conditions imply that those activities which are used at a positive level in the equilibrium solution make a profit of zero.

Given the assumption of continuous and single-valued excess demand functions and the description of the production possibility set by means of an activity analysis model, the following rather direct application of Brouwer's Theorem is sufficient to demonstrate the existence of a equilibrium solution. Under weaker assumptions on the model, variants such as Kakutani's Fixed-Point Theorem may be required.

Computation of General Equilibria, Fig. 3



Let prices be normalized so as to lie on the unit simplex $S = \{p = (p_1,p_2,\ldots,p_n) | p_i \geq 0, \sum pi = 1\}$. The set of prices p for which $pA \leq 0$ is termed the *dual* cone of the production possibility set generated by the activity matrix A. Its intersection with the unit simplex is a convex polyhedron C consisting of those normalized prices which yield a profit less than or equal to zero for all activities.

We construct a continuous mapping of S into itself as follows: for each p in S consider the point $p + f(p)$; a point which is generally not on the unit simplex itself. We then define $g(p)$ – the image of p under the mapping – to be that point in C which is closest, in the sense of Euclidean distance, to $p + f(p)$. It is then an elementary application of the Kuhn–Tucker Theorem to show that a fixed point of this mapping is, indeed, an equilibrium price vector.

The Equilibrium Model as a Tool for Policy Evaluation

Brouwer's original proof of his theorem was not only difficult mathematically, but it was decidedly non-constructive; it offered no method for effectively computing a fixed point of the mapping. Brouwer did, in fact, reject his own argument during the later 'intuitionist' phase of his career, in which he proclaimed the acceptability of only those mathematical conclusions obtained by constructive procedures. In spite of the many simplifications in the proof of Brouwer's Theorem offered during the subsequent half-century, it was not until the mid-1960s that constructive methods for approximating fixed points of a continuous mapping finally made their appearance on the scene (Scarf 1967) – aided by the development of the modern electronic computer and by the rapid methodological advances in the discipline of operations research.

In the early decades of this century, the question of the explicit numerical solution of the general equilibrium model was an active topic of discussion – not by numerical analysts – but rather by economists concerned with the techniques of economic planning in a socialist economy. The issue was raised in the remarkable paper published by Enrico Barone in 1908, entitled 'The Ministry of Production in a Socialist Economy'. Barone, and subsequently Oskar Lange (1936), accepted the Walrasian model – with suitable transfers of income – as an adequate description of ideal economic activity in an economy in which the means of production were collectively owned. In the absence of markets, prices, levels of output and the choice of productive techniques were to be obtained by an explicit numerical solution of the Walrasian system. A key feature of Barone's analysis was the concept of the 'technical coefficients of production' – the input/output coefficients associated with those activities in use at equilibrium. Barone's contention was that the equilibrium could be found – by an extremely laborious calculation which might indeed claim a significant share of the national product – only if the correct activities were known in advance. For Barone, rational economic calculation in a socialist economy was defeated by the many opportunities for substitution in production: the particular activities in use at equilibrium would be impossible to determine by a prior computation. It is instructive to quote Barone on this point.

> The determination of the coefficients economically most advantageous can only be done in an *experimental* way: and not on a *small scale*, as could be done in a laboratory; but with experiments on a *very large scale*, because often the advantage of the variation has its origin precisely in a new and greater dimension of the undertaking. Experiments may be successful in the sense that they may lead to a lower cost combination of factors; or they may be unsuccessful, in which case the particular organization may not be copied and repeated and others will be preferred, which *experimentally* have given a better result.
>
> The Ministry of Production could not do without these experiments for the determination of the *economically* most advantageous technical coefficients if it would realize the condition of the minimum cost of production which is *essential* for the attainment of the maximum collective welfare.
>
> It is on this account that the equations of the equilibrium with the maximum collective welfare are not soluble *a priori*, on paper.

An Elementary Algorithm

Barone's negative conclusion is certainly valid if the full production possibility set, including all of

the possibilities for substitution in production, is not known to the central planner. In this event, numerical calculation is impossible, and Lange's suggestion, made some 20 years later, may be appropriate: the problem can be turned on its head and the market, itself, can be used as a mechanism of discovery as well as a giant analogue computer. But if the production possibility set can be explicitly constructed, substitution – in and of itself – does not seem to me to be a severe impediment to numerical computation.

At the present moment, some 20 years after the introduction and continued refinement of fixed-point computational techniques, I have in my possession a small floppy disk with a computer program which will routinely solve – on a personal computer – for equilibrium prices and activity levels in a Walrasian model in which the number of variables is on the order of 100. (The authors of the program suggest that examples with 300 variables can be accommodated on a mainframe computer.) Substantial possibilities of substitution, if known in advance, offer no difficulty to the successful functioning of this algorithm. In my opinion, the modern restatement of Barone's problem is rather that even 300 variables are extremely small in number in contrast to the millions of prices and activity levels implicit in his account. The computer, while expanding our capabilities immeasurably, has taught us a severe lesson about the role of mathematical reasoning in economic practice and forced us to shift our point of view dramatically from that held by our predecessors. We realize that our preoccupations are not with universal laws which describe economic phenomena with full and complete generality, but rather with intellectual formulations which are an imperfect representation of a complex and elusive reality. The application of general equilibrium theory to economic planning, and more generally to the evaluation of the consequences of changes in economic policy, must be based on highly aggregated models whose conclusions are at best tentative guides to action.

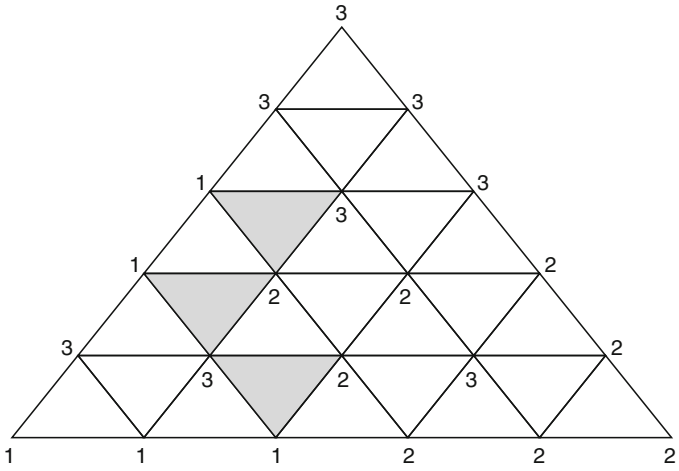
An exercise in comparative statics is begun by constructing a general equilibrium model whose solution reflects the economic situation existing prior to the proposed policy change. The number of parameters required to describe demand functions, initial endowments and the production possibility set is considerable, and in practice the constraint of reproducing the current equilibrium must be augumented by a variety of additional statistical estimates in order to specify the model. The limitations of data in the form required by the Walrasian model inevitably make this estimation procedure less than fully satisfactory.

The second step in the exercise is to calculate the solution after the proposed policy changes are explicitly introduced into the model. In some cases the policy variables being studied can be directly incorporated as parameters in the equations whose solution yields the equilibrium values; if the changes are small, their effects on the solution may be obtained by differentiating these equations and solving the resulting linear system for the corresponding changes in the equilibrium values themselves. This approach was adopted by Leif Johansen (1960) and by Arnold Harberger (1962) in his study of the incidence of a tax on corporate profits. The use of this method in policy analysis continues in Norway, and it forms the basis of the amibitious programme carried out by Peter Dixon and his collaborators in Australia (1982). If, on the other hand, the policy changes are large, the equilibrium position may be shifted substantially, and its determination may require the use of more sophisticated computational methods.

Fixed-point algorithms can be divided into two major classes: those based on the elements of differential topology, surveyed by Smale (1981), and those which are combinatorial in nature. The most elementary of the combinatorial algorithms for approximating a fixed point of a continuous mapping of the unitsimplex $S = \{(x = (x_1, x_2, \ldots, x_n) | x_i \geq 0, \sum x_i = 1\}$ begins by dividing the simplex into a large number of small subsimplices as illustrated in Fig. 4. In our notation the simplex is of dimension $n-1$ and has faces of dimension $n-2, \ldots, 1$. It is a requirement of the subdivision that the intersection of any two of the subsimplices is either empty or a full lower dimensional face of both of them.

Each vertex of the subdivision will have associated with it an integer label selected from the set

Computation of General Equilibria, Fig. 4



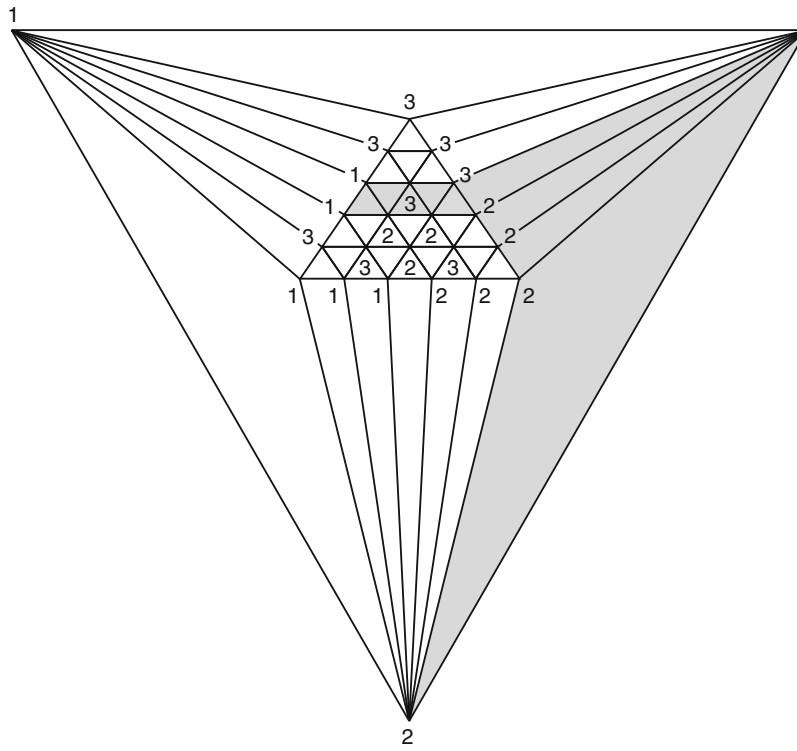
(1, 2,. . ., *n*). When the method is applied to the determination of a fixed point of a particular mapping, the labels associated with a vertex will depend on the mapping evaluated at that point. For the moment, however, the association will be arbitrary aside from the requirement that a vertex on the boundary of the simplex will have a label *i* only if the *i*th coordinate of that vertex is positive.

The remarkable combinatorial lemma demonstrated by Emanuel Sperner (1928) in his doctoral thesis is that at least one subsimplex must have all of its vertices differently labelled. Assuming this result to be correct, let us consider a mapping of the simplex in which the image of the vector $x = (x_1,\ldots,xn)$ is $f(x) = [f_1(x),\ldots, f_n(x)]$. The requirement that the image be on the simplex implies that $f_i(x) \geq 0$ and that $\sum fi(x) = 1$. It follows that for every vertex of the subdivision *v*, unless *v* is a fixed point of the mapping, there will be a least one index *i* for which $fi(v) - vi < 0$. If we select such an index to be the label associated with the vertex *v*, then the assumptions of Sperner's Lemma are clearly satisfied, and the conclusion asserts the existence of a simplex whose vertices are distinctly labelled.

If the simplicial subdivision is very fine, the vertices of this sub-simplex are all close together; at each vertex a different coordinate is decreasing under the mapping, and by continuity every point in the small subsimplex will have the property that each coordinate is not increasing very much under the mapping. Since the sum of the coordinate changes is by definition zero, the image of any point in the completely labelled subsimplex will be close to itself, and such a point will therefore serve as an approximate fixed point of the mapping. A formal proof of Brouwer's Theorem requires us to construct a sequence of finer and finer subdivisions, to find, for each subdivision, a completely labelled simplex, and to select a convergent sequence of these simplices tending to a fixed point of the mapping.

Sperner's Lemma may be applied to the equilibrium problem directly. For simplicity, consider the model of exchange in which the market excess demand functions are given by $g(p)$, with p on the unit price simplex. As before, we subdivide the simplex and associate an integer label from the set (1,. . ., *n*) with each vertex *v* of the subdivision, according to the following rule: the label *i* is to be selected from the set of those indices of which $g_i(p) \leq 0$. It is an elementary consequence of Walras's Law that a selection can be made which is consistent with the assumptions of Sperner's Lemma, and there will therefore be a subsimplex all of whose vertices bear distinct labels. By virtue of the particular labelling rule, any point in such a completely labelled simplex will be an approximate equilibrium price vector in the sense that all excess demands, at this price, will be either negative or, if positive, very small.

Computation of General Equilibria, Fig. 5



Sperner's original proof of his combinatorial lemma was not constructive; it was based on an inductive argument which required a complete enumeration of all completely labelled simplices for a series of lower dimensional problems. In order to develop an effective numerical algorithm for the determination of such a simplex let us begin by embedding the unit simplex, and its subsimplices, in a larger simplex T, as in Fig. 5. The larger simplex is subdivided by joining its n new vertices to those vertices of the original subdivision lying on the boundary of the unit simplex. The assumptions of Sperner's Lemma permit the new vertices to be given distinct labels from the set $(1,\ldots, n)$, in such a way that no additional completely labelled simplices are generated. For concreteness, let the new vertex receiving the label i be denoted by v^i.

We begin our search for a completely labelled simplex by considering the simplex with vertices $v^2,\ldots, v^n$ and one additional vertex, say v^*. If v^* has the label 1, this simplex is completely labelled and our search terminates; otherwise we move to an adjacent simplex by removing the vertex whose label agrees with that of v^* and replacing it with that unique other vertex yielding a simplex in the subdivision. As the process continues, we are, at each step, at a simplex whose vertices bear the labels $2,\ldots,n$, with a single one of these labels appearing on a pair of vertices. Precisely two $n-2$ dimensional faces have a complete set of labels $2,\ldots,n$. The simplex has been entered through one of these faces; the algorithm proceeds by exiting through the other such face.

The argument first introduced by Lemke (1965) in his study of two person non- zero sum games was carried over by Scarf (1967) to show that the above algorithm never returns to a simplex previously visited and never requires a move outside of T. Since the number of simplices is finite, the algorithm must terminate, and termination can only occur when a completely labelled simplex is reached.

Improvements in the Algorithm

The algorithm can easily be programmed for a computer, and it provides the most elementary numerical procedure for approximating fixed points of a continuous mapping and equilibrium prices for the Walrasian model. Since its introduction in 1967, the algorithm, in this particular form, has been applied to a great number of examples of moderate size, and it performs sufficiently well in practice to conclude that the numerical determination of equilibrium prices is a feasible undertaking. The algorithm does, however, have some obvious drawbacks which must be overcome to make it available for problems of significant size. For example, the information which yields the labelling of the vertices, and therefore the path taken by the algorithm, is simply the index of a coordinate which happens to be decreasing when the mapping is evaluated at the vertex. More recent algorithms make use of the full set of coordinates of the image of the vertex instead of a single summary statistic.

Second, this primitive algorithm is always initiated at the boundary of the simplex. If the approximation is not sufficiently good, the grid size must be refined, and a recalculation, which makes no use of previous information, must be performed. It is of the greatest importance to be able to initiate the algorithm at an arbitrary interior point of the simplex selected as our best a priori estimate of the answer.

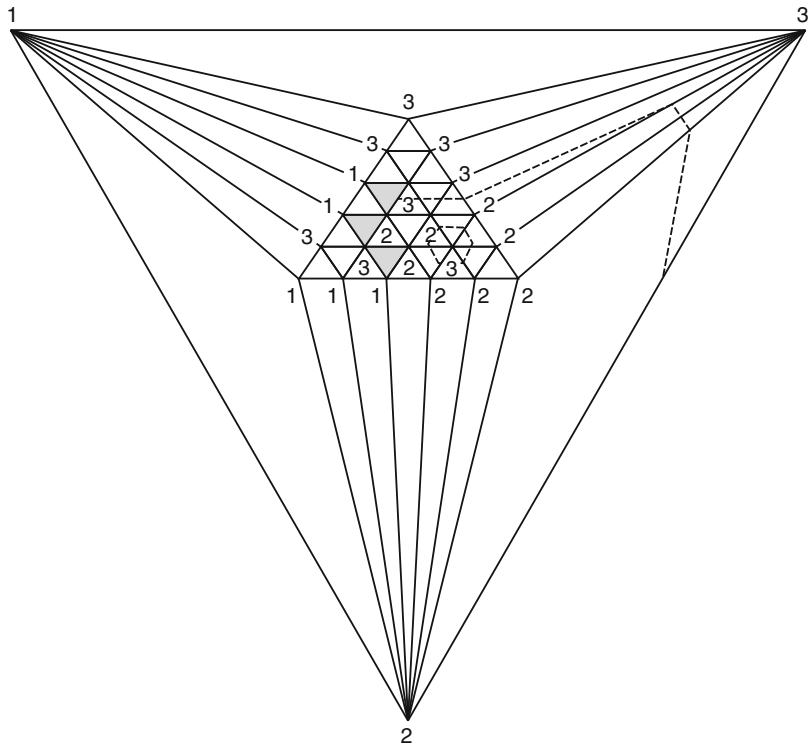
The following geometrical setting (Eaves and Scarf 1976) for the elementary algorithm suggests the form these improvements can take. Let us construct a piecewise linear mapping, $h(x)$, of T into itself as follows: for each vertex v in the subdivision let $h(v)=v^i$, where i is the label associated with v. We then complete the mapping by requiring h to be linear in each simplex of the subdivision. The mapping is clearly continuous on T and maps every boundary point of T into itself. Moreover, every subsimplex in the subdivision whose vertices are not completely labelled is mapped, by h, into the boundary of T. If none of the simplices were completely labelled, this construction would yield a most improbable conclusion: a continuous mapping of T into itself which is the identity on the boundary and which maps the entire simplex into the boundary. That such a mapping cannot exist is known as the Non-Retraction Theorem, an assertion which is, in fact, equivalent to Brouwer's Theorem. The impossibility of such a mapping reinforces our conclusion that a completely labelled simplex does exist.

Select a point c interior to one of the boundary faces of T and consider the set of points which map into c; that is, the set of x for which $h(x)=c$. As Fig. 6 indicates, this set contains a piecewise linear path beginning at the point c, and transversing precisely those simplices encountered in our elementary algorithm. There are however, other parts of the set $\{x|h(x)=c\}$: closed loops which do not touch the boundary of T and other piecewise linear paths connecting a pair of completely labelled simplices. Stated somewhat informally, the general conclusion, of which this is an example, is that the inverse image of a particular point, under a piecewise linear mapping from an n dimensional set to an $n-1$ dimensional set, consists of a finite union of interior loops, and paths which join two boundary points (see Milnor 1965, for the differentiable version).

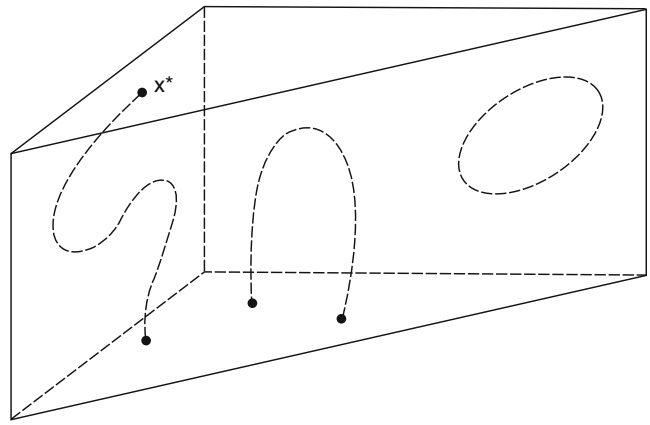
To see how this observation can be used, consider the product of the unit simplex S and the closed unit interval [0, 1]; that is, the set of points (x, t) with x in S and $0 \leq t \leq 1$, as in Fig. 7. Extend the mapping from the unit simplex to this large set by defining $F(x,t)=(1-t)\,f(x)+tx^*$, with x^* a preselected point on the simplex, taken to be an estimate of the true fixed point. The set of points for which $F(x,t)-x=0$ is, by our general conclusion, a finite union of paths and loops. Precisely one of these paths intersects the upper boundary of the enlarged set. If the path is followed, its other endpoint must lie in the face $t=0$ and yield a fixed point of the original mapping.

The path leading to the fixed point can be followed on the computer in several ways. We can, for example, introduce a simplicial decomposition of the set $S \times [0, 1]$ and approximate F by a piecewise linear mapping agreeing with F on the vertices of the subdivision. Following the path then involves the same type of calculation we have become accustomed to in carrying out linear programming pivot steps. There are a great many variations in the mode of simplicial subdivision

Computation of General Equilibria, Fig. 6



Computation of General Equilibria, Fig. 7



leading to substantial improvements in the efficiency of our original fixed-point algorithm (Eaves 1972; Merrill 1971; van der Laan and Talman 1979).

An alternative procedure, adopted by Kellogg et al. (1976) and Smale (1976), is to impose sufficient regularity conditions on the underlying mapping so that differentiation of $F(x,t)-x = 0$ yields a set of differential equations for the path joining x^* to the fixed point on $t = 0$. This leads to a variant of Newton's method which is global in the sense that it need not be initiated in the vicinity of the correct answer. But, whichever of these alternatives we select, the numerical

difficulties in computing equilibrium prices can be overcome for all problems of reasonable size.

Applied General Equilibrium Analysis

During the last 15 years, the field of Applied General Equilibrium Analysis has grown considerably; instead of the few tentative examples illustrating our ability to solve general equilibrium problems, we have seen the construction of a large number of models of substantial size designed to illuminate specific policy issues. The number of books and papers which have appeared in the field is far too large for a complete enumeration in this essay, and I shall mention only a few public ations which may be consulted to obtain an indication of the diversity of this activity. The paper by Shoven and Whalley (1984) in the *Journal of Economic Literature* is a survey of applied general equilibrium models in the fields of taxation and international trade constructed by these authors and their colleagues. The volume by Adelman and Robinson (1978) is concerned with the application of general equilibrium analysis to problems of economic development. Whalley (1985) has written on trade liberalization, and Ballard et al. (1985) on the evaluation of tax policy. Jorgenson (Hudson and Jorgenson 1974) and Manne (1976) have made extensive applications of this methodology to energy policy, and Ginsburg and Waelbroeck (1981) provide a refreshing discussion of alternative computational procedures applied to a model of international trade involving over 200 commodities. The volume edited by Scarf and Shoven (1985) contains a collection of papers presented at one of an annual series of workshops in which both applied and theoretical topics of interest to researchers in the field of Applied General Equilibrium Analysis are discussed.

See Also

- ▶ Computation of General Equilibria (New Developments)
- ▶ General Equilibrium

Bibliography

Adelman, I., and S. Robinson. 1978. *Income distribution policy in developing countries: A case study of Korea*. Stanford: Stanford University Press.

Ballard, C.L., D. Fullerton, J.B. Shoven, and J. Whalley. 1985. *A general equilibrium model for tax policy evaluation*. Chicago: University of Chicago Press.

Barone, E. 1908. Il Ministerio della Produzione nello stato colletivista. *Giornale degli Economisti e Revista di Statistica*. Trans. as The Ministry of Production in the Collectivist State. In *Collectivist economic planning*, ed. F.A. Hayek. London: G. Routledge & Sons, 1935.

Brouwer, L.E.J. 1912. ÜberAbbildungen von Mannigfaltigkeiten. *Mathematische Annalen* 71: 97–115.

Debreu, G. 1982. Existence of competitive equilibrium. In *Handbook of mathematical economics*, ed. K.J. Arrow and M. Intriligator. Amsterdam: North- Holland.

Dixon, P.B., B.R. Parmenter, J. Sutton, and D.P. Vincent. 1982. *ORANI: A multisectoral model of the Australian economy*. Amsterdam: North-Holland.

Eaves, B.C. 1972. Homotopies for the computation of fixed points. *Mathematical Programming* 3: 1–22.

Eaves, B.C., and H. Scarf. 1976. The solution of systems of piecewise linear equations. *Mathematics of Operations Research* 1: 1–27.

Ginsburg, V.A., and J.L. Waelbroeck. 1981. *Activity analysis and general equilibrium modelling*. Amsterdam: North-Holland.

Harberger, A. 1962. The incidence of the corporation income tax. *Journal of Political Economics* 70: 215–240.

Hudson, E.A., and D.W. Jorgenson. 1974. US energy policy and economicgrowth. *Bell Journal of Economics and Management Science* 5: 461–514.

Johansen, L. 1960. *A multi-sectoral study of economic growth*. Amsterdam: North-Holland.

Kellogg, R.B., T.Y. Li, and J. Yorke. 1976. A constructive proof of the Brouwer fixed point theorem and computational results. *SIAM Journal of Numerical Analysis* 13: 473–483.

Kuhn, H.W. 1968. Simplicial approximation of fixed points. *Proceedings of the National Academy of Sciences* 61: 1238–1242.

Lange, O. 1936. On the economic theory of socialism. *Review of Economic Studies* 4 (53–71): 123–142.

Lemke, C.E. 1965. Bimatrix equilibrium points and mathematical programming. *Management Science* 11: 681–689.

Manne, A.S. 1976. ETA: A model of energy technology assessment. *Bell Journal of Economics and Management Science* 7: 379–2406.

Merrill, O.H. 1971. Applications and extensions of an algorithm that computers fixed points of certain nonempty convex upper semicontinuous point to set mappings. Technical Report 71–7, University of Michigan.

Milnor, J. 1965. *Topology from the differentiable viewpoint*. Charlottesville: University of Virginia Press.

C

Scarf, H.E. 1967. The approximation of fixed points of a continuous mapping. *SIAM Journal of Applied Mathematics* 15: 1328–1343.
Scarf, H.E., with the collaboration of T. Hansen. 1973. *The computation of economic equilibria*. London/New Haven: Yale University Press.
Scarf, H., and J.B. Shoven, eds. 1984. *Applied general equilibrium analysis*. Cambridge: Cambridge University Press.
Shoven, J.B., and J. Whalley. 1972. A general equilibrium calculation of the effects of differential taxation of income from capital in the U.S. *Journal of Public Economy* 1: 281–321.
Shoven, J.B., and J. Whalley. 1984. Applied general-equilibrium models of taxation and international trade. *Journal of Economic Literature* 22: 1007–1051.
Smale, S. 1976. A convergent process of price adjustment and global Newton methods. *Journal of Mathematical Economics* 3: 107–120.
Smale, S. 1981. Global analysis and economics. In *Handbook of mathematical economics*, ed. K.J. Arrow and M. Intriligator, vol. I. Amsterdam: North-Holland.
Sperner, E. 1928. Neur Beweis für die Invarianz der Dimensionszahl und des Gebietes. *Abhandlungen an den mathematischen Seminar der Universität Hamburg* 6: 265–272.
van der Laan, G., and A.J.J. Talman. 1979. A restart algorithm for computing fixed points without an extra dimension. *Mathematical Programming* 17: 74–84.
Whalley, J. 1985. *Trade liberalization among major world trading areas*. Cambridge, MA: MIT Press.

Computation of General Equilibria (New Developments)

Felix Kubler

Abstract

In this article, I review two recent developments in the theory of computation of general equilibria. First, following Brown et al. (1996) several papers have developed globally convergent algorithms for the computation of general equilibria in models with incomplete asset markets. I review some of the developments in that area. Second, new developments in computational algebraic geometry lead to algorithms to compute effectively all equilibria of systems of polynomial equations. I point out some applications of these algorithms to general equilibrium theory.

Keywords

Computation of general equilibria; Gröbner bases; Homotopy algorithms; Incomplete asset markets; Kuhn–Tucker conditions; Multiple equilibria; Newton–Kantarovich conditions; Real business cycles; Semi-algebraic economies; Smale's alpha method; Tarski–Seidenberg th; Uncertainty

JEL Classifications

D5

Introduction

After Scarf (1967) showed that there exist globally convergent (and effectively applicable) algorithms to compute economic equilibria, there is now a class of computable applied models which are routinely used to evaluate the economic consequences of different taxes and tariff structures (see, for example, Shoven and Whalley 1992). Research on efficient algorithms for the computation of general equilibria in these models largely took place outside of economics.

A large literature in numerical analysis has developed algorithms that are much faster than Scarf's original method and that can be used for large-scale applications. Efficient iterative schemes, mostly based on global Newton methods, now allow applied researchers to solve for competitive equilibria in models with hundreds of commodities and agents (see, for example, Ferris and Pang 1997).

Recently, there has been substantial research in theoretical computer science on the development of polynomial time algorithms for the computation of general equilibria. For most existing methods, the number of operations needed to approximate equilibria within a fixed precision ε grows exponentially in $1/\varepsilon$. Under restrictive assumptions on preferences, in models without production, researchers have developed algorithms to approximate equilibria 'in polynomial time', that is, the running time of the algorithm increases polynomially in the input parameters and in the precision with which equilibria are

computed. Codenotti et al. (2004) give an overview on recent developments along this line.

In this article I will not discuss any of these practical aspects of the solution of large-scale models. I will instead focus on the following two unrelated developments in the computation of general equilibria in economics.

1. The computation of equilibria in models with time, uncertainty and missing asset markets.
2. The computation of all equilibria and the relationship between exact and approximate equilibria in the standard Arrow–Debreu model.

Models with Asset Markets

Due to their essential static nature, standard computable general equilibrium models suffer from an oversimplified treatment of uncertainty. Agents either solve a static problem or have myopic expectations, and the model can therefore not explicitly incorporate investment and saving decisions. The general equilibrium model with incomplete asset markets (GEI model) provides a basic framework with several agents and several commodities to incorporate uncertainty and financial markets. See, for example, Magill and Qunizii (1996) for an overview of the literature. The computation of equilibria in these models is challenging because in some specifications equilibria fail to exist while in others they are often numerically unstable.

Kehoe and Prescott (1995) argue that real business cycle models provide an alternative way to extend computable general equilibrium to models with time and uncertainty. There is now a large literature on the computation of equilibria in dynamic stochastic economies. This is reviewed elsewhere in this dictionary; see approximate solutions to dynamic models (linear methods); see also Judd (1998).

In the standard GEI model there are two time periods (Kubler and Schmedders 2000, show how the problem of computation of equilibria in multi-period finance models can be essentially reduced to the two period case) and S possible states of the world in the second period. There are L perishable commodities available for trade at each state. There are H agents with endowments $e^h \in \mathbb{R}_+^{(S+1)L}$ and utility functions $u^h : \mathbb{R}_+^{(S+1)L} \to \mathbb{R}$. It is assumed throughout this article that utility functions are smooth in the sense of Debreu (1972) – that is, utility is C^2, strictly increasing, strictly quasi-concave, exhibits non-zero Gaussian curvature and indifference curves do not cut the axes.

There are J assets available for trade. In each state s, asset j pays a bundle of commodities $a_j(s) \in \mathbb{R}^L$. It is without loss of generality to assume that the $LS \times J$ matrix

$$A = \begin{pmatrix} a_1(1) & \dots & a_J(1) \\ \vdots & \ddots & \vdots \\ a_1(S) & \dots & a_J(S) \end{pmatrix}$$

has full rank J. Allowing assets to pay in different commodities is crucial when one wants to extend the model to several time periods and long-lived securities.

In the following, it will be useful to write commodity prices as

$$\begin{aligned} p &= (p(0), p(1), \dots, p(S)) \in \Delta^{(S+1)L-1} \\ &= \left\{ p \in \mathbb{R}_+^{(S+1)L} : \sum_i p_i = 1 \right\}, \end{aligned}$$

and the $S \times J$ asset payoff matrix (as a function of spot prices $p(1) \dots p(S)$), $R(p)$, as

$$R(p) = \begin{pmatrix} p(1) \cdot a_1(1) & \dots & p(1) \cdot a_J(1) \\ \vdots & \ddots & \vdots \\ p(S) \cdot a_1(S) & \dots & p(S) \cdot a_J(S) \end{pmatrix}.$$

In part of the discussion we assume an exogenous short-sale constraint, that is, there is a number $0 < K \leq \infty$ such that the two-norm of an agent's portfolio must always be less than or equal to K. One can then write an agent's aggregate excess demand function as the solution of his maximization problem in the GEI economy.

$$\begin{aligned} \left(z^h(p), \varphi^h(p)\right) &= \arg \max_{z \in \mathbb{R}^{L(S+1)}, \varphi \in \mathbb{R}^J} u\left(e^h + z\right) \text{ s.t.} p \cdot z \\ &= 0(p(1) \cdot z(1), \dots, p(S) \cdot z(S))^T \\ &= R(p) \cdot \varphi \| \varphi \| \leq K. \end{aligned}$$

A GEI equilibrium is a collection of prices, portfolios and a consumption allocation such that markets clear and each agent maximizes her utility, i.e. equilibrium prices p are characterized by $\sum_{h=1}^{H} z^h(p) = 0$.

In a slight idealization (see also the more precise definition in the next section), we assume that the maximization problem can be solved exactly and we define an ε-equilibrium as a price $\overline{p}$ such that

$$||\sum_{h=1}^{H} z^h(\overline{p})|| < \varepsilon.$$

A General Algorithm

Although generally $R(p)$ will have full rank J, there will be so-called 'bad prices' at which the rank of $R(p)$ drops. When there are no short sale constraints, that is, $K = \infty$, this leads to a discontinuity of excess demand. Scarf's algorithm fails: no matter how fine the simplicial subdivision, if the algorithm terminates at some $\overline{p}$, one cannot necessarily infer a bound on $\|z(\overline{p})\|$ and hence cannot find an ε-equilibrium.

Homotopy continuation methods (see Garcia and Zangwill 1981; Eaves 1972) turn out to be ideally suited for this numerical problem. In order to solve a system of equations $f(x) = 0, f: X \to Y$, the basic idea underlying homotopy methods is to find a smooth map $H : X \times [0, 1] \to Y$ with

$$H(x, 1) \equiv f(x) \text{ and } H(x, 0) \equiv g(x),$$

where $g : X \to Y$ has a known unique zero. The map H is called a smooth homotopy. In using homotopy methods it is crucial to set up the function, H, to ensure that there is a smooth path that connects $(x^s,0)$ with $g(x^s) = 0$ to some $(\overline{x}, 1)$ with $f(\overline{x}) = 1$.

Brown et al. (1996) develop a homotopy algorithm which can be shown to be globally convergent in that it finds an ε-equilibrium for any $\varepsilon > 0$ in a finite number of steps. Following the so-called Cass-trick, it is useful to introduce an unconstrained agent, that is, to define the first agent maximization problem as

$$z^u(p) = \arg\max_z u^1(e + z) \text{ s.t.} p \cdot z = 0,$$

and aggregate demand as $z(p) = z^u(p) + \sum_{h=2}^{H} z^h(p)$. Note that $\overline{p}$ is a GEI equilibrium (given that $K = \infty$) if and only if $z(p) = 0$. An ε-equilibrium is characterized by $\|z(p)\| < \varepsilon$.

Define the expenditure of the unconstrained agent y^u as

$$y^u = \big(p(1) \cdot z_1^u(p), \ldots, p(S) \cdot z_S^u(p)\big).$$

Define an extended payoff matrix $R^*(p)$ by

$$R * (p) = [R(p), y^u(p)]$$

and let $R^*_{-i}(p)$ be $R^*(p)$ with the i'th column deleted. For the constrained agents $h = 2, \ldots, H$ define

$$z^h\big(p, R^*_{-i}(p)\big) = \arg\max_{z,\varphi} u^h\big(e^h + z\big) \text{ s.t.} p \cdot z = 0$$

$$(p(1) \cdot z(1), \ldots, p(S) \cdot z(S))^T = R^*_{-i}(p) \cdot \varphi.$$

Now consider a family of homotopies, indexed by i

$$H_i(p, t, \theta) = \begin{pmatrix} z^u(p) + t\sum_{h=2}^{H} z^h\big(p, R^*_{-i}(p)\big) \\ R * (p)\theta \\ \theta \cdot \theta - 1 \end{pmatrix}.$$

To prove existence of a homotopy path, Brown et al. (1996) show that $\cup_{i=1}^{J+1} H_i^{-1}$ contains a smooth path connecting the starting point to a solution at $t = 1$.

While generically in endowments a homotopy path turns out to exist, the algorithm is hardly applicable in medium-sized problems, since the number of homotopies one has to consider can become quite large. An alternative is to focus on models with $K < \infty$ (or alternatively models with transaction costs) or to consider algorithms which might fail in a small class of problems but which are generally more efficient.

Short-Sale Constraints

In the presence of short-sale constraints, the excess demand function is continuous and equilibrium existence can be proven with Brouwer's theorem. Therefore, one could presumably use a version of Scarf's algorithm to compute equilibria in this case. However, while there are no new mathematical problems to be solved, the fact that the rank of the asset–payoff matrix can still collapse in equilibrium poses difficult numerical problems. Simple Newton method-based algorithms often do not work (see Kubler and Schmedders 2000) unless one has a starting point very close to the actual solution. It turns out that, just as in the problem without short-sale constraints, homotopy continuation methods can provide a basis for reliable algorithms.

Schmedders (1998) develops a homotopy algorithm which can be used to solve models with a large number of heterogeneous households and goods. The basic idea of his algorithm is to modify the agents' problem by introducing a homotopy parameter $t \in [0, 1]$ as follows.

$$\left(z^h(p,t),\varphi^h(p,t)\right) = \arg \max_{z\in\mathbb{R}^{L(S+1)},\varphi\in\mathbb{R}^J} u\left(e^h+z\right) - (1-t)\frac{1}{2}||\varphi||^2 \text{ s.t. } p\cdot z = 0(p(1)\cdot z(1),\ldots,p(S)\cdot z(S)) = R(p)\cdot\varphi||\varphi|| \le K.$$

Under the assumptions on utilities this is still a convex problem and the first order Kuhn–Tucker conditions are necessary and sufficient. Schmedders provides various examples that show that even for $K = \infty$ his algorithm, although not guaranteed to converge, performs well in practice.

For $K < \infty$, the Kuhn–Tucker inequalities can be converted into a system of equalities via a change of variables (see Garcia and Zangwill 1981, ch. 4). Kubler (2001), Herings and Schmedders (2006) and others subsequently used this idea to solve models with transaction costs, trading constraints and other market imperfections.

Of course, it is an important practical problem how to trace out a homotopy path numerically. See Watson (1979) for a theoretical algorithm. For a practical description of numerical homotopy path-following methods see Schmedders (2004).

Equilibria in Semi-algebraic Economies

While it is clear that sufficient assumptions for the global uniqueness of competitive equilibria are too restrictive to be applicable to models used in practice, it remains an open problem how serious a challenge the non-uniqueness of competitive equilibrium poses to applied equilibrium modelling. In the presence of multiple equilibria, comparative statics exercises become meaningless. Furthermore, even when for a given specification of the economy equilibrium is globally unique, as Richter and Wong (1999) point out, the possibility of multiple equilibria for close-by economies implies that it is generally impossible to compute prices and allocations that are close-by exact equilibrium prices and allocations (as opposed to computing prices at which aggregate excess demand is close to zero). In this section I argue that one can solve these problems by focusing on so-called 'semi-algebraic' economies.

While the arguments are also applicable to the GEI model, for simplicity, consider a standard Arrow–Debreu exchange economy, $\left(u^h, e^h\right)_{h=1}^H$. There are H agents trading L commodities. Each agent h has individual endowments $e^h \in \mathbb{R}^L_+$ and 'smooth preferences' characterized by an utility function $u^h : \mathbb{R}^L_+ \to \mathbb{R}$.

A Walrasian equilibrium is a collection of consumption vectors $\left(x^h\right)_{h=1}^H$ and prices $p \in \Delta^{L-1}$ such that

$$x^h \in \arg\max_{x\in\mathbb{R}^L_+} u^h(x) \text{ s.t.} p\cdot x \le p\cdot e^h \qquad (1)$$

$$\sum_{h=1}^H \left(x^h - e^h\right) = 0. \qquad (2)$$

An approximate (ε-) equilibrium consists of an allocation an prices such that

$$||u^h(x^h) - \left[\max_{x \in \mathbb{R}^L_+} u^h(x) \text{ s.t.} p \cdot x \leq p \cdot e^h\right]|| < \varepsilon \quad (3)$$

$$||\sum_{h=1}^{H} (x^h - e^h)|| < \varepsilon. \quad (4)$$

Given any $\varepsilon > 0$, Scarf's algorithm (as well as the more efficient algorithms used in practice) finds a p, x^h which constitute an ε-equilibrium.

This leaves open two important theoretical questions.

1. Can one relate the approximate equilibrium prices and allocations, to exact equilibria, that is, given a computed ε-equilibrium $(\overline{p}, (\overline{x}^h))$, does there exist a Walrasian equilibrium $\tilde{p}, (\tilde{x}^h)$ with $\|(\overline{p}, (\overline{x}^h)) - (\tilde{p}, (\tilde{x}^h))\|$ small? Can one find good bounds on this distance which tend to zero as $\varepsilon \to 0$?
2. Given an economy $(u^h, e^h)_{h=1}^H$ with N Walrasian equilibria $(p^h, (x^h)^n)_{n=1}^N$ and any $\delta > 0$, is it possible to approximate all N equilibria, that is, to find N ε-equilibria $(\tilde{p}^n, (\tilde{x}^h)^n)_{n=1}^N$ with $\|(p^n, (x^h)^n) - (\tilde{p}^n, (\tilde{x}^h)^n)\| < \delta$, for all $n = 1, \ldots, N$?

Clearly, the second problem is strictly more difficult to tackle than the first. Richter and Wong (1999) show that for general economies even the answer to the first question is negative. In order to obtain positive answers to both qsts, one needs to restrict possible preferences. One approach is to assume that better sets are semialgebraic sets. I will make the slightly more useful assumption that marginal utilities are semi-algebraic functions.

Semi-algebraic Economies

We assume that for each h, $D_x u^h(x)$ is a semi-algebraic function, that is, its graph $\{(x, y) \in \mathbb{R}^{2L}_+ : y = D_x u^h(x)\}$ is a finite union and intersection of sets of the form

$$\{(x, y) \in \mathbb{R}^{2L} : g(x, y) > 0\} \text{ or}$$
$$\{(x, y) \in \mathbb{R}^{2L} : f(x, y) = 0\}$$

for polynomials with real coefficients, f and g.

For practical purposes, the focus on semi-algebraic preferences is quite general. First note that Afriat's theorem implies that a finite set of observations on an individual's choices that can be rationalized by any utility function can also be rationalized by semi-algebraic preferences (in fact, Afriat's construction is piece-wise linear). Furthermore, note that the constant elasticity of substitution utility function which is often used in applied work is semi-algebraic if the elasticities of substitution are rational numbers.

It follows from the Tarski–Seidenberg theorem that for semi-algebraic economies the answers to both qsts above are positive, since the relevant statements can be written as first order sentences (see Basu et al. 2003). However, algorithmic quantifier elimination which needs to be used to answer general qsts in this framework is so computationally inefficient that for practical purposes this does not help towards solving the above qsts for interesting specifications of economies.

Nevertheless, given a semi-algebraic economy it is possible to find a system of polynomial equations $f(x) = 0, f: \mathbb{R}^{H(L+1)+L-1} \to \mathbb{R}^{H(L+1)+L-1}$, and finitely many inequalities $g^i(x) \geq 0$, $g^i : \mathbb{R}^{H(L+1)+L-1} \to \mathbb{R}^M$, $i = 1, \ldots, N < \infty$ such that p, (x^h) is a Walrasian equilibrium for the economy (u^h, e^h) if and only if there exist $\lambda^h \in \mathbb{R}_{++}$, $h = 1, \ldots, H$ such that for some $i = 1, \ldots, N$,

$$f(p, (x^h, \lambda^h)) = 0, g^i(p, (x^h, \lambda^h)) \geq 0.$$

Therefore, the problem of finding Walrasian equilibria reduces to finding the real roots of polynomial systems of equations and verifying polynomial inequalities (see Kubler and Schmedders 2006).

Having reduced the problem of finding Walrasian equilibria to finding roots of a polynomial system of equations, one can then answer the two qsts above affirmatively.

Question 1: Smale's Alpha Method

Smale's alpha method provides a simple sufficient conditions for approximate zeros to be close to exact zeros and can be viewed as an extension of the Newton–Kantarovich conditions. The following results are from Blum et al. (1998, ch. 8).

Let $D \subset \mathbb{R}^n$ be open and let $f : D \to \mathbb{R}^n$ be analytic. For $z \in D$, define $f^{(k)}(z)$ to be the k'th derivative of f at z. This is a multi-linear operator which maps k-tuples of vectors in D into $\mathbb{R}^n$. Define the norm of an operator A to be

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|}.$$

Suppose that the Jacobian of f at z, $f^{(1)}(z)$ is invertible and define

$$\gamma(z) = \sup_{k \geq 2} \left\| \frac{\left(f^{(1)}(z)\right)^{-1} f^{(k)}(z)}{k!} \right\|^{\frac{1}{(k-1)}}$$

and

$$\beta(z) = \left\| \left(f^{(1)}(z)\right)^{-1} f(z) \right\|.$$

Theorem 1 Given a $\overline{z} \in D$, suppose the ball of radius $\left(1 - \frac{\sqrt{2}}{2}\right)/\gamma(\overline{z})$ around $\overline{z}$ is contained in D and that

$$\beta(\overline{z})\gamma(\overline{z}) < 0.157.$$

Then there exists a $\tilde{z} \in D$ with

$$f(\tilde{z}) = 0 \text{ and } \|\overline{z} - \tilde{z}\| \leq 2\beta(\overline{z}).$$

While the theorem applies to any locally analytic function, the bound $\gamma(z)$ can in general only be obtained if the system is in fact polynomial. For this case, the bound can be computed fairly easily. Given an ε-equilibrium the result gives an immediate bound on the distance between the approximation and an exact Walrasian equilibrium, hence answering Question 1 above.

Question 2: Polynomial System Solving

In the following, I denote the collection of all polynomials in the variable $x_1, x_2, \ldots, x_n$ with coefficients in a field K by K $[x_1, \ldots, x_n]$. The for this survey relevant examples of K are the field of rational numbers $\mathbb{Q}$, the field of real numbers $\mathbb{R}$, and the field of complex numbers $\mathbb{C}$. Polynomials over the field of rational numbers are computationally convenient since modern computer algebra systems perform exact computations over the field $\mathbb{Q}$. Economic parameters are typically real numbers, and equations characterizing equilibria lie in $\mathbb{R}[x]$. The algorithms to compute all solutions to polynomial systems always compute all solutions in an algebraically closed field, in this case $\mathbb{C}[x]$.

Given a polynomial system of equations $f: \mathbb{C}^M \to \mathbb{C}^M$ there is now a variety of algorithm to approximate numerically all complex and real zeros of f. Sturmfels's monograph (2002) provides an excellent overview. In this survey I briefly mention two possible approaches, homotopy continuation methods and solution methods based on Gröbner bases.

At the writing of this article, both approaches are too inefficient to be applicable to large economic models, but they can be used for models with four or five households and four or five commodities. To find all equilibria for a given economy, homotopy methods seem slightly more efficient, while Gröbner bases allow for statements about entire classes of economies.

All Solution Homotopies

Solving polynomial systems numerically means computing approximations to all isolated solutions. Homotopy continuation methods can provide paths to all approximate solutions. There are well-known bounds on the maximal number of complex solutions of a polynomial system. The basic idea is to start at a generic polynomial system $g(x)$ whose number of roots is at least as large as the maximal number of solutions to $f(x) = 0$ and whose roots are all known. Then one needs to trace out all paths (in complex space) of the homotopy $H(x, t) = tg(x) + (1 - t)f(x)$, which do not diverge to infinity. Smale's alpha method can be applied along the path to ensure that the approximate solutions are close to real exact solutions (see Blum et al. 1998). It can be shown that all solutions to $f(x) = 0$ can be found in this manner.

Sommese and Wampler (2005) provide a detailed overview. Applications of these methods in economics have so far been largely restricted to game theory, but the method is also applicable to Walrasian equilibria.

Gröbner Basis

For given polynomials $f_1, \ldots, f_k$ in $\mathbb{Q}[x]$ the set

$$I = \left\{ \sum_{i=1}^{k} h_i f_i : h_i \in \mathbb{Q}[x] \right\} = \langle f_1, \ldots, f_k \rangle$$

is called the ideal generated by $f_1, \ldots, f_k$. It turns out that under conditions which can often be shown to hold in practice, the so-called 'reduced Gröbner basis' of this ideal, I, in the lexicographic term order has the shape

$$\mathcal{G} = \{x_1 - q_1(x_n), x_2 - q_2(x_n), \ldots, x_{n-1} - q_{n-1}(x_n), r(x_n)\}$$

where r is a polynomial of degree d and the q_i are polynomials of degree $d - 1$.

This basis can be computed exactly, using Buchberger's algorithm (recently, much more efficient versions of the basic algorithm have been developed; see for example Faugère 1999). The number of real solutions to the original system then equals the number of real solutions of the univariate polynomial $r(.)$ which can be determined exactly by Sturm's method (see Sturmfels 2002, for details). The roots of $r(.)$ can be approximated numerically with standard methods and the remaining solution to the original system is linear in these roots.

Kubler and Schmedders (2006) use the method to test for uniqueness of equilibria in semi-algebraic classes of economies.

See Also

- ▶ Approximate Solutions to Dynamic Models (Linear Methods)
- ▶ Computation of General Equilibria
- ▶ General Equilibrium

Bibliography

Basu, S., R. Pollack, and M.-F. Roy. 2003. *Algorithms in real algebraic geometry*. New York: Springer.

Blum, L., F. Cucker, M. Shub, and S. Smale. 1998. *Complexity and real computation*. New York: Springer.

Brown, D.J., P.M. DeMarzo, and B.C. Eaves. 1996. Computing equilibria when asset markets are incomplete. *Econometrica* 64: 1–27.

Codenotti, B., S. Pemmaraju, and K. Varadarajan. 2004. Algorithms column: The computation of market equilibria. *ACM SIGACT News* 35(4): 23–37.

Debreu, G. 1972. Smooth preferences. *Econometrica* 40: 603–615.

Eaves, B.C. 1972. Homotopies for the computation of fixed points. *Mathematical Programming* 3: 1–22.

Faugère, J.C. 1999. A new efficient algorithm for computing Gröbner bases (f4). *Journal of Pure and Applied Algebra* 139: 61–88.

Ferris, M.C., and J.S. Pang. 1997. Engineering and economic applications of complementarity problems. *SIAM Review* 39: 669–713.

Garcia, C., and W. Zangwill. 1981. *Pathways to solutions, fixed points, and equilibria*. Englewood Cliffs: Prentice Hall.

Herings, P.J.J., and K. Schmedders. 2006. Computing equilibria in finance economies with incomplete markets and transaction costs. *Economic Theory* 27: 493–512.

Judd, K. 1998. *Numerical methods in economics*. Cambridge: MIT Press.

Kehoe, T.J., and E.C. Prescott. 1995. Introduction to the symposium, the discipline of applied general equilibrium. *Economic Theory* 6: 1–11.

Kubler, F. 2001. Computable general equilibrium with financial markets. *Economic Theory* 18: 73–96.

Kubler, F., and K. Schmedders. 2000. Computing equilibria in stochastic finance economies. *Computational Economics* 15: 145–172.

Kubler, F., and Schmedders, K. 2006. Uniqueness of equilibria in semi-algebraic economies. Discussion paper, Northwestern University.

Magill, M.J.P., and M. Qunizii. 1996. *Theory of incomplete markets*. Cambridge: MIT Press.

Richter, M.K., and K.-C. Wong. 1999. Non-computability of competitive equilibrium. *Economic Theory* 14: 1–27.

Scarf, H. 1967. On the computation of equilibrium prices. In *Ten economic studies in the tradition of irving fisher*, ed. W.J. Fellner. New York: Wiley.

Schmedders, K. 1998. Computing equilibria in the general equilibrium model with incomplete asset markets. *Journal of Economic Dynamics and Control* 22: 1375–1403.

Schmedders, K. 2004. Homotopy path-following with easyhomotopy: Solving nonlinear equations for economic models. Working paper, Northwestern University.

Shoven, J.B., and J. Whalley. 1992. *Applying general equilibrium*. Cambridge: Cambridge University Press.

Sommese, A.J., and C.W. Wampler. 2005. *The numerical solution of systems of polynomials arising in engineering and science*. Singapore: World Scientific Press.

Sturmfels, B. 2002. *Solving systems of polynomial equations*, CBMS Regional Conference Series in

Mathematics No. 97. Providence: American Mathematical Society.
Watson, L.T. 1979. A globally convergent algorithm for computing fixed points of C2 maps. *Applied Mathematics and Computation* 5: 297–311.

Computational Methods in Econometrics

Vassilis A. Hajivassiliou

Abstract

The computational properties of an econometric method are fundamental determinants of its importance and practical usefulness, in conjunction with the method's statistical properties. Computational methods in econometrics are advanced through successfully combining ideas and methods in econometric theory, computer science, numerical analysis, and applied mathematics. The leading classes of computational methods particularly useful for econometrics are matrix computation, numerical optimization, sorting, numerical approximation and integration, and computer simulation. A computational approach that holds considerable promise for econometrics is parallel computation, either on a single computer with multiple processors, or on separate computers networked in an intranet or over the internet.

Keywords

Bayesian inference; Bootstrap; Classical inference; Computational methods; Generalized least squares; Generalized method of moments; Importance sampling simulation; Jackknife; Least absolute deviations; Maximum likelihood; Numerical integration; Optimal control; Ordinary least squares; Random effects models; Simulation-based estimation; Stone, J. R. N; Markov chain Monte Carlo methods; Parallel computation

JEL Classification

C15

Introduction

In evaluating the importance and usefulness of particular econometric methods, it is customary to focus on the set of *statistical* properties that a method possesses – for example, unbiasedness, consistency, efficiency, asymptotic normality, and so on. It is crucial to stress, however, that meaningful comparisons cannot be completed without paying attention also to a method's *computational* properties. Indeed the practical value of an econometric method can be assessed only by examining the inevitable interplay between the two classes of properties, since a method with excellent statistical properties may be computationally infeasible and vice versa. Computational methods in econometrics are evolving over time to reflect the current technological boundaries as defined by available computer hardware and software capabilities at a particular period, and hence are inextricably linked with determining what the state of the art is in econometric methodology.

To give a brief illustration, roughly from the late 1950s until the early 1960s we had the 'Stone Age' of econometrics, when the most sophisticated computational instrument was the slide rule, which used two rulers on a logarithmic scale, one sliding into the other, to execute approximate multiplication and division. In this Stone Age, suitably named in honour of Sir Richard Stone, winner of the 1984 Nobel Prize in Economics, the brightest Ph.D. students at the University of Cambridge were toiling for days and days in back rooms using slide rules to calculate ordinary linear regressions, a task which nowadays can be achieved in a split second on modern personal computers.

The classic linear regression problem serves to illustrate the crucial interaction between statistical and computational considerations in comparing competing econometric methods. Given data of size S, with observations on a dependent variable denoted by $S \times 1$ vector y and corresponding observations on k explanatory factors denoted by $S \times k$ matrix X ($k < X$), the linear plane fitting exercise is defined by Gauss's minimum quadratic distance problem:

$$\begin{aligned}\widehat{\beta} = & \arg\min_b (y - Xb)'(y - Xb) \\ \times \equiv & \arg\min_b \sum_{s=1}^{S} \left(y_s - x_s' b\right)^2 \end{aligned} \quad (1)$$

where x_s' is the sth row of matrix X and b is a $k \times 1$ vector of real numbers defining the regression plane Xb. Under the assumption that X has full column rank k, the solution to this *ordinary least squares* minimization problem is the linear-in-y expression $\widehat{\beta} = (X'X)^{-1}X'y$, which only requires the matrix operations of multiplication and inversion. Suppose, however, that Gauss had chosen instead as his measure of distance the sum of absolute value of the deviations, and defined instead:

$$\tilde{\beta} = \text{rg} \min_b \sum_{s=1}^{S} |y_s - x_s' b| \quad (2)$$

The vector $\tilde{\beta}$ that solves the second minimization is known as the *least absolute deviations* (LAD) estimator and has no closed-form matrix expression. In fact, calculation of $\tilde{\beta}$ requires highly nonlinear operations for which computationally efficient algorithms were developed only in the 1970s. To give a concrete example, consider the *intercept-only* linear regression model where X is the $S \times 1$ vector of ones. Then the single $\widehat{\beta}$ coefficient that solves (1) is the sample mean of y, while $\tilde{\beta}$ that solves (2) is the sample *median* of y. The latter is orders of magnitude more difficult to compute than the former since it involves sorting y and finding the value in the middle, while the former simply adds all elements of y and divides by the sample size. Clearly, it could be quite misleading if $\widehat{\beta}$ and $\tilde{\beta}$ where compared solely in terms of statistical properties without any consideration of their substantially different computational requirements.

A second example in a similar vein is the following parametric estimation problem. Suppose a sample of size S is observed on a single variable y. It is believed that each observation y_s is drawn independently from the same uniform distribution on the interval $[\theta,c]$ where the lower value of the support is the single unknown parameter that needs to be estimated, while c is known. Two parametric estimation methods with particularly attractive statistical properties are the generalized method of moments (GMM) and the method of maximum likelihood (MLE). Indeed, for relatively large sample sizes these two methods are comparably attractive in terms of statistical properties, while they differ *drastically* in terms of computational requirements: the GMM solution is $\widehat{\theta}_{gmm} = \frac{2}{S}\sum_{s=1}^{S} y_s - c$, thus requiring only the simple calculation of the sample mean y, while the MLE involves the highly nonlinear operation of finding the minimum of the data vector y, $\widehat{\theta}_{mle} = min(y_1, \ldots, y_S)$.

In the following section we discuss in turn the leading classes of methods that are of particular importance in modern econometrics, while section "Parallel computation" introduces the concept of parallel processing and describes its current value and future promise in aiding dramatically econometric computation.

Computational Methods Important for Econometrics

The advancement of computational methods for econometrics relies on understanding the interplay between the disciplines of econometric theory, computer science, numerical analysis, and applied mathematics. In the five subsections below we discuss the leading classes of computational methods that have proven of great value to modern econometrics.

Matrix Computation and Specialized Languages

To start with the fundamental econometric framework of linear regression, the *sine qua non* of econometric computation is the ability to program and perform efficiently matrix operations. To this end, specialized matrix computer languages have

been developed which include Gauss and Matlab. Fundamental estimators of the linear regression coefficient vector β, like the OLS $(X'X)^{-1}X'y$ and its generalized least squares (GLS) variant $(X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y$, are leading examples of the usefulness of such matrix languages, where the $S \times S$ matrix Ω is a positive definite, symmetric variance-covariance matrix of the disturbance vector $\varepsilon \equiv y - X\beta$. Matrix operations are useful even for nonlinear econometric methods discussed below, since a generally useful approach is to apply linearization approximations through the use of differentiation and Taylor's expansions.

In implementing econometric methods that involve matrix operations, special attention needs to be paid to the dimensionality of the various matrices, as well as to any special properties a matrix may posses, which can affect very substantially the feasibility and performance of the computational method to be adopted. Looking at the OLS and GLS formulae, we see three different matrices that require inversion: $X'X$, Ω, and $X'\Omega^{-1}X$. The first and the third are of dimension $k \times k$, while the second is $S \times S$. Since the number of regressors k is typically considerably smaller than the sample size S, the inversion of these matrices can involve vastly different burden in terms of total number of computer operations required as well as memory locations necessary for holding the information during those calculations. (For example, in panel data settings where multiple observations are observed in different time-periods for a cross-section of economic agents, it is not uncommon to have total sample sizes of 300,000 or more.) To this end, econometric analysts have focused on importing from numerical analysis matrix algorithms that are particularly efficient in handling sparse as opposed to dense matrices. By their very nature, sparse matrices exhibit a very high degree of compressibility and concomitantly lower memory requirements. See Drud (1977) for the use of sparse matrix techniques in econometrics. A matrix is called sparse if it is primarily populated by zeros, for example, the variance-covariance matrix of a disturbance vector following the moving-average-of-order-1 model:

$$\Omega_{ma1} = \sigma^2 \begin{pmatrix} 1 & \frac{\lambda}{1+\lambda^2} & 0 & \cdots & 0 \\ \frac{\lambda}{1+\lambda^2} & 1 & \frac{\lambda}{1+\lambda^2} & \ddots & \vdots \\ 0 & \frac{\lambda}{1+\lambda^2} & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 1 & \frac{\lambda}{1+\lambda^2} \\ 0 & \cdots & 0 & \frac{\lambda}{1+\lambda^2} & 1 \end{pmatrix}$$

In contrast, a stationary autoregressive disturbance of order 1 has a dense variance-covariance matrix:

$$\Omega_{ar1} = \sigma^2 \begin{pmatrix} 1 & \gamma & \gamma^2 & \cdots & \gamma^{S-1} \\ \gamma & 1 & \gamma & \ddots & \vdots \\ \gamma^2 & \gamma & \ddots & \ddots & \gamma^2 \\ \vdots & \ddots & \ddots & 1 & \gamma \\ \gamma^{S-1} & \cdots & \gamma^2 & \gamma & 1 \end{pmatrix}$$

Other matrix algebra methods especially important in econometrics are the Cholesky factorization (see Golub 1969) of a positive definite matrix A into the product $A = R'R$ where R is an upper-triangular matrix, and the singular value decomposition that allows the calculation of pseudo-inverse of any matrix B which may be non-square, and if square, not positive definite (see Belsley 1974).

It is important to note that on occasion a brilliant theoretical development can simplify enormously the computational burden of econometric methods that, though possessing attractive statistical properties, were thought to be infeasible with existing computation technology in the absence of the theoretical development. A case in point is the GLS/MLE estimator for the one-factor random effects model proposed by Balestra and Nerlove (1966), which is of great importance in the analysis of linear panel data models. The standard formulation gives rise to the GLS formula

requiring the inversion of an equi-correlated variance covariance matrix Ω of dimension $S \times S$, where S is of the order of the product of the number of available observations in the cross-section dimension times the number available in the time dimension. For modern panel data-sets, this can exceed 300,000, thus making the calculation of Ω^{-1} infeasible even on today's supercomputers, let alone with the slide rules available in 1966. Fuller and Battese (1973), however, showed that the equi-correlated nature of the one-factor random effects model made calculation of the GLS estimator equivalent to an OLS problem, where the dependent variable $\tilde{y}$ and the regressors $\tilde{X}$ are simple linear combinations of the original data $y_{it}, x_{1it}, \ldots, x_{kit}$ and its time averages $\bar{y}_{it}, \bar{x}_{1it}, \ldots, \bar{x}_{kit}$ defined by $\bar{y}_{i.} \equiv \frac{1}{T}\sum_{t=1}^{T} y_{it}$ and $\tilde{y}_{it} \equiv y_{it} - \lambda \bar{y}_{i.}$, and analogously for the regressor variables. This realization allowed the calculation of the GLS estimator without the need for inverting the usually problematically large Ω matrix.

Another important case where a theoretical development in methodology led to a dramatic lowering of the computational burden and hence allowed the calculation of models that would otherwise have had to wait perhaps for decades for sufficient advancements in computer technology is the simulation-based inference for Limited Dependent Variable models, associated with the name of Daniel McFadden (1989). See section "Computer simulation" below, McFadden, Daniel and simulation-based estimation.

Optimization

Many econometric estimators with attractive statistical properties require the optimization of a (generally) nonlinear function of the form:

$$q \equiv \arg \max_{\theta} F(\theta; data) \tag{3}$$

over a vector of unknown parameters θ of dimension p, typically considerably larger than 1. Examples are: the method of maximum likelihood, minimum-distance (OLS, LAD, GMM), and other extremum estimators. (The need to optimize functions numerically is also important for certain problems in computational economics, for example, the problem of optimal control.) Algorithms for optimizing functions of many variables are a key component in the collection of tools for econometric computation. The suitability of a certain algorithm to a specific optimization econometric problem depends on the following classification:

1. *Algorithms that require the calculation of first and possibly second derivatives* Versus *algorithms that do not*. Clearly, if the function to be optimized is not twice continuously differentiable (as is the case with LAD) or even discontinuous (as is the case with the maximum score estimator for the semiparametric analysis of the binary response model – see Manski 1975), algorithms that require differentiability will not be suitable. The leading example of an algorithm not relying on derivatives is the nonlinear simplex method of Nelder and Meade (1965).
2. *Local* Versus *global algorithms*. Optimization algorithms of the first type (for example, Gauss-Newton, Newton-Raphson, and Berndt et al. (1974)) search for an optimum in the vicinity of the starting values fed into the algorithm. This strategy may not necessarily lead to a global optimum over the full set of parameter space. This is of particular importance if the function to be optimized has multiple local optima, where typically the estimator with the desirable statistical properties corresponds to locating the overall optimum of the function. In such cases, global optimization algorithms (for example, simulated annealing and genetic optimization algorithm) should be employed instead.

Special methods are necessary for constrained optimization, where a function must be maximized or minimized subject to a set of equality or inequality constraints. These problems, in general considerably more demanding than unconstrained optimization, can be handled through three main alternative approaches: interior, exterior and re-parameterization methods.

Comprehensive reviews of optimization methods in econometrics can be found in Goldfeld

and Quandt (1972), Quandt (1983), and Dennis and Schnabel (1984). These studies also discuss the related issue of the numerical approximation of derivatives and illustrate the fundamental link in terms of computation between optimization and the problem of solving linear and nonlinear equations. For similar methods used in economics, see numerical optimization methods in economics and nonlinear programming.

Sorting

Of special importance for computing the class of estimators known as robust or semiparametric methods is the ability to sort data rapidly and computationally efficiently. Such a need arises in the calculation of order statistics, for example, the sample median and sample minimum required by the first two estimation examples given above. The leading sorting algorithms, bubble-, heap- and quick-sort, have fundamentally different properties in terms of computation speed and memory requirements, in general depending on how close to being sorted the original data series happens to be. For a practical review of the leading sorting algorithms, see Press et al. (2001, ch. 8).

Numerical Approximation and Integration

Numerical approximation is necessary for any mathematical function that does not have a closed form solution, for example, exponential, natural logarithm and error functions. See Abramowitz and Stegun (1964) for an exhaustive study of mathematical functions and their efficient approximation. Judd (1996) focuses on numerical approximation methods particularly useful in economics and econometrics.

Numerical integration, also known as numerical quadrature, is a related approximation problem that is crucial to modern econometrics. There are two key fields of econometrics where integrals without a closed form must be evaluated numerically. The first is Bayesian inference where moments of posterior densities need to be evaluated, which take the form of high-dimensional integrals. See, inter alia, Zellner et al. (1988). The second main class is classical inference in limited dependent variable (LDV) models; for example, Hajivassiliou and Ruud (1994). See Geweke (1996) for an exhaustive review of numerical integration methods in computational economics and econometrics, and Davis and Rabinowitz (1984) for earlier results.

It is important to highlight a crucial difference between the numerical integration problems in Bayesian inference and those in classical inference for LDV models, which makes various integration-by-simulation algorithms be useful to one field and not the other: in the Bayesian case, typically a single or a few high-dimensional integrals have to be evaluated accurately. In contrast, in the classical LDV inference case, quite frequently hundreds of thousands of such integrals need to be approximated.

Computer Simulation

The need for efficient generation of pseudo-random numbers with good statistical properties on a computer appears very routinely in econometrics. Leading examples include:

- Statistical methods based on resampling, primarily the 'jackknife' and the 'bootstrap', as introduced by Efron (1982). These methods have proven of special value in improving the small sample properties of certain econometric estimators and test procedures, for example in reducing estimation bias. They are also used to approximate the small sample variance of estimators for which no closed form expressions can be derived.
- Evaluation of econometric estimators through Monte Carlo experiments, where hypothetical data-sets with certain characteristics are simulated repeatedly and the econometric estimators under study are calculated for each set. This allows the calculation of empirical (simulated) properties of the estimators, either to compare to theoretical mathematical calculations or because the latter are intractable.
- Calculation of frequency probabilities of possible outcomes in large-scale decision trees, for which the outcome probabilities are impossible to characterize theoretically.
- Sensitivity analyses and what-if studies, where an econometric model is 'run' on a computer under different scenarios of policy measures.

- Simulation-based Bayesian and classical inference, where integrals are approximated through computer simulation (known as Monte Carlo integration). Particularly important methods in this context are the following: frequency simulation; importance sampling; and Markov chain Monte Carlo methods (the leading exponents being Gibbs resampling and the Metropolis/Hastings algorithm). A related class of methods, known as variance-reduction simulation techniques, includes control variates and antithetics. See Geweke (1988) and Hajivassiliou et al. (1996) for reviews. See also simulation-based estimation.

Parallel Computation

Parallel processing, where a computation task is broken up and distributed across different computers, is a technique that can afford huge savings in terms of total time required for solving particularly difficult econometric problems. For example, the simulation-based estimators mentioned in the previous section exhibit the potential of significant computational benefits by calculating them on computers with massively parallel architectures, because the necessary calculations can be organized in essentially an independent pattern. An example of such a computer is the Connection Machine CM-5 at the National Center for Supercomputing Applications in Illinois with 1024 identical processors in a multiple-instruction/multiple-data (MIMDI) configuration. The benefits of such a parallel architecture on the problem of solving an econometric optimization classical estimator not involving simulation can also be substantial, since such estimators involve the evaluation of contributions to the criterion (for example, likelihood) function in the case of independently and identically distributed (i.i.d.) observations. Since typical applications in modern applied econometrics using cross-sectional and longitudinal data sets involve several thousands of i.i.d. observations, the potential benefits of parallel calculations of such estimators should be obvious. The benefits of a massively parallel computer architecture become even more pronounced in the case of simulation-based estimators. See Nagurney (1996) for a discussion of parallel computation in econometrics.

An alternative approach for parallel computation that does not involve a single computer with many processors has been developed recently and offers considerable promise for computational econometrics. Through the use of specialized computer languages, many separate computers are harnessed together over an organization's intranet or even over the internet, and an econometric computation task is distributed across them. The benefits of this approach depend critically on the relative burden of the overhead of communicating across the individual computers when organizing the splitting of the tasks and then collecting and processing the separate partial results. Such distributed parallel computation has the exciting potential of affording formidable super-computing powers to econometric researchers with only modest computer hardware.

See Also

- ▶ Longitudinal Data Analysis
- ▶ McFadden, Daniel (Born 1937)
- ▶ Non-linear Programming
- ▶ Numerical Optimization Methods in Economics
- ▶ Robust Estimators in Econometrics
- ▶ Simulation-Based Estimation

Bibliography

Abramowitz, M., and I. Stegun. 1964. *Handbook of mathematical functions*. Washington, DC: National Bureau of Standards.

Balestra, P., and M. Nerlove. 1966. Pooling cross-section and time-series data in the estimation of a dynamic model. *Econometrica* 34: 585–612.

Belsley, D. 1974. Estimation of system of simultaneous equations and computational specifications of GREMLIN. *Annals of Economic and Social Measurement* 3: 551–614.

Berndt, E.K., B.H. Hall, R.E. Hall, and J.A. Hausman. 1974. Estimation and inference in nonlinear structural models. *Annals of Economic and Social Measurement* 3: 653–666.

Davis, P.J., and P. Rabinovitz. 1984. *Methods of numerical integration*. New York: Academic.
Dennis, J.E., and R.B. Schnabel. 1984. *Unconstrained optimization and nonlinear equations*. Englewood Cliffs: Prentice-Hall.
Drud, A. 1977. An optimization code for nonlinear econometric models based on sparse matrix techniques and reduced grades. *Annals of Economic and Social Measurement* 6: 563–580.
Efron, B. 1982. *The jackknife, the bootstrap, and other resampling plans*, CBMS-NSF monographs No. 38. Philadelphia: SIAM.
Fuller, W.A., and G.E. Battese. 1973. Transformations for estimation of linear models with nested-error structure. *Journal of the American Statistical Association* 68: 626–632.
Geweke, J. 1988. Antithetic acceleration of Monte Carlo integration in Bayesian inference. *Journal of Econometrics* 38: 73–90.
Geweke, J. 1996. Monte Carlo simulation and numerical integration. In *Handbook of computational economics*, vol. 1, ed. H. Amman, D. Kendrik, and J. Rust. Amsterdam: North-Holland.
Goldfeld, S., and R. Quandt. 1972. *Nonlinear methods in econometrics*. Amsterdam: North-Holland.
Golub, G.H. 1969. Matrix decompositions and statistical calculations. In *Statistical computation*, ed. R.C. Milton and J.A. Milder. New York: Academic.
Hajivassiliou, V.A., and P.A. Ruud. 1994. Classical estimation methods using simulation. In *Handbook of econometrics*, vol. 4, ed. R. Engle and D. McFadden. Amsterdam: North-Holland.
Hajivassiliou, V.A., D.L. McFadden, and P.A. Ruud. 1996. Simulation of multivariate normal rectangle probabilities and derivatives: Theoretical and computational results. *Journal of Econometrics* 72(1, 2): 85–134.
Judd, K. 1996. Approximation, perturbation, and projection methods in economic analysis. In *Handbook of computational economics*, vol. 1, ed. H. Amman, D. Kendrik, and J. Rust. Amsterdam: North-Holland.
Manski, C. 1975. Maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics* 3: 205–228.
McFadden, D. 1989. A method of simulated moments for estimation of multinomial discrete response models. *Econometrica* 57: 995–1026.
Nagurney, A. 1996. Parallel computation. In *Handbook of computational economics*, vol. 1, ed. H. Amman, D. Kendrik, and J. Rust. Amsterdam: North-Holland.
Nelder, J.A., and R. Meade. 1965. A simplex method for function minimization. *Computer Journal* 7: 308–313.
Press, W.H., B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling. 2001. *Numerical recipes in Fortran 77: The art of scientific computing*. Cambridge: Cambridge University Press.
Quandt, R. 1983. Computational problems and methods. In *Handbook of econometrics*, vol. 1, ed. Z. Griliches and M. Intriligator. Amsterdam: North-Holland.
Zellner, A., L. Bauwens, and H. VanDijk. 1988. Bayesian specification analysis and estimation of simultaneous equation models using Monte Carlo methods. *Journal of Econometrics* 38: 73–90.

C

Computer Industry

Shane Greenstein

Abstract

Commercial computing has grown to include an extraordinary range of economic undertakings. In any given era, computing markets are organized around platforms – a cluster of technically standardized components that buyers use together to make the wide range of applications. There has been an increasing secular trend in the number of firms that possess the necessary technical knowledge and commercial capabilities to bring to market some component or service. While general improvements in technical capabilities are readily apparent, it is quite difficult to calculate the productivity improvements arising from increased investment in and use of computing.

Keywords

Computer industry; Economic growth; Information technology; Innovation; Moore's law

JEL Classifications

L63

The commercial computing industry accounts for a large fraction of economic activity. From its military and research origins in the late 1940s, it spread into the commercial realm and has since grown to include an extraordinary range of economic undertakings. Many economists believe this expansion of applications for computing has been a driver of economic growth.

Computing aids the automated tracking of transactions, a function that finds use, for

example, in automating billing, managing the pricing of inventories of airline seating, and restocking retail outlets in a geographically dispersed organization. It also facilitates the coordination of information-intensive tasks, such as the dispatching of time-sensitive deliveries or emergency services. Computing also enables performance of advanced mathematical calculations, useful in such diverse activities as calculating interest on loans and generating estimates of underground geological deposits. Computer-aided precision also improves the efficiency of processes such as manufacturing metal shapes or the automation of communication switches, to name just two.

In any given era, computing markets are organized around platforms – a cluster of technically standardized components that buyers use together to make the aforementioned wide range of applications. Such platforms involve long-lived assets, both components sold in markets (that is, hardware and some software) and components made by buyers (that is, training and most software). Important computing platforms historically include the UNIVAC, the IBM 360 and its descendents, the Wang minicomputers, IBM AS/400, DEC VAX, Sun SPARC, Intel/Windows PC, Unix/Linux, and, after the mid-1990s, TCP/IP-based client–server platforms linked together.

Vendors tend to sell groups of compatible products under umbrella strategies aimed at the users of particular platforms. In the earliest eras of computing markets, the leading firms integrated all facets of computing and offered a supply of goods and services from a centralized source. In later eras, the largest and most popular platforms historically included many different computing, communications and peripheral equipment firms, software tool developers, application software writers, consultants, system integrators, distributors, user groups, news publications and service providers.

Until the early 1990s, most market segments were distinguished by the size of the tasks to be undertaken and by the technical sophistication of the typical user. Mainframes, minicomputers, workstations, and personal computers, in decreasing order, constituted different size-based market segments. Trained engineers or programmers made up the technical user base, while the commercial market was geared more towards administrators, secretaries and office assistants.

The most popular platform in the late 1980s and 1990s differed from the prominent platforms of earlier years. The personal computer (PC) began in the mid- 1970s as an object of curiosity among technically skilled hobbyists, but became a common office tool after the entry of IBM's design. Unlike prior computing platforms, this one has diffused into both home and business use. From the beginning, this platform involved thousands of large and small software developers, third-party peripheral equipment and card developers, and a few major players. In more recent experience, control over the standard has completely passed from IBM to Microsoft and Intel. Microsoft produces the Windows operating system and Intel produces the most commonly used microprocessor. For this reason the platform is often called Wintel.

The networking and internet revolution in the late 1990s is responsible for blurring once-familiar distinctions. These new technologies have made it feasible to build client–server systems within large enterprises and across ownership boundaries. It employs internet-based computing systems networked across potentially vast geographic distances, supporting the emergence of a 'network of networks'.

Despite frequent and sometimes dramatic technical improvements in specific areas of technology, many features of the most common platforms in use tend to persist or change very slowly. Many durable components make up platforms. And, though they lose their market value as they become obsolete in comparison with frontier products, they do not as quickly lose their ability to provide a flow of services to users. Consequently, new technology tends to be most successful when new components enhance and preserve the value of previous investments, a factor that creates demand for 'backward compatible' upgrades or improvements. It also creates a demand for support and service activities to reduce the costs of making the transition from old to new.

Control over changes to design and other aspects of technical standards shapes the backward compatibility for key components. Control of these decisions is coincident with platform leadership – determining the rate and direction of change in technical features of components around which other firms build their businesses. In each platform, it is very rare to observe more than a small number of firms acquiring leadership positions. Since such positions have been historically associated with high firm profitability, firms compete fiercely for market dominance in component categories where standards are essential. Not surprisingly, competitive behaviour affiliated with obtaining and retaining market leadership does occasionally receive attention from antitrust authorities.

Though innovative change in computing began well prior to the invention of the integrated circuit, in popular discussion advances in computing have become almost synonymous with advances in microprocessors. This is due to an observation by Gordon Moore, who co-founded and became chairman at Intel. In 1965 he foresaw a doubling of circuits per chip every two years. This prediction about the rate of technical advance later became known as 'Moore's law'. In fact, microprocessors and DRAMS have been doubling in capability every 18 months since the mid-1970s.

Moore's prediction pertained narrowly to integrated circuits. However, a similar pattern of improvement – though with variation in the rate – characterizes other electronic components that go into producing a computer or that are complementary with computing in many standard uses. This holds for disk drives, display screens, routing equipment, and data-transmission capacity, to name a few. Such widespread innovation creates opportunities for new entry and rearrangements in the conditions of supply.

Accordingly, there has been an increasing secular trend in the number of firms that possess the necessary technical knowledge and commercial capabilities to bring to market some component or service of value to computing users. This factor alone explains the increasing complexity of supply chains for the supply of most computing hardware and software products. It is also coincident with their increasing geographic reach. In addition, as in other manufacturing processes, the increasing use of sophisticated information technology helps coordinate design and production involving firms from many countries and continents.

While the spawning of new information technology businesses in North America has tended to be concentrated in a small number of locations, such as the Boston area and Santa Clara Valley (popularly known as Silicon Valley), every other facet of the supply chain for computing involves firms headquartered and operating in a much wider set of locations. In North America, these range from Seattle, Austin, Los Angeles, the greater New York area, Denver–Boulder, Washington DC, the North Carolina Research Triangle, Chicago, and virtually all major cities in the United States. The supply chain for many complementary components has also been associated with many firms in Western Europe and as well as in India, Israel, South Korea, Singapore, Taiwan and China. Even more widespread are computing service firms, which follow business and home users dispersed across the globe.

Despite this geographic dispersion since the 1950s, US companies have retained leadership in generating new platforms and commercializing frontier technologies in forms that most users find valuable. Part of this results from the persistence of platform leadership for a time within a segment. In addition, US firms have historically been ascendant whenever platform leadership has changed. However, this pattern seems likely to change in the 21st century, as non-US firms already have found leadership positions in producing components of many platforms and in related areas of electronics, such as consumer electronics, communication equipment and specialized software.

While general improvements in technical capabilities are readily apparent, it is quite difficult to calculate the productivity improvements arising from increased investment in and use of computing. There is no question that existing computing activities have become less expensive, while new capabilities have been achieved. This has allowed

economic actors to attain previously unobtainable outcomes. This shift in economic possibilities has generated a restructuring of organizational routines, market relationships, and other activities associated with the flow of goods, which inevitably improves the economy's ability to transform inputs into consumer welfare.

Yet altering the business use of computing can be slow. It often demands large adjustment costs and gradual learning about which organizational processes can best employ advances in computing. It can involve a reallocation of decision rights and discretion inside a large organization, especially when business units alter a wide array of intermediate routine processes (such as billing, account monitoring, and inventory management) or the coordination of services (such as the delivery of data for decision support). Moreover, the largest changes come from altering many complementary activities that respond to new and unanticipated opportunities, setting off new waves of invention. Each wave's productivity effect is interwoven with others.

Along with these improvements the boundaries of the 'computing market' have changed. A hardware-based definition for the computing market was barely adequate in the 1960s and is no longer adequate for economic analysis. However, there is no consensus about what alternative framing will be appropriate for understanding value creation, supplier behaviour, and user adoption in computing in the 21st century.

See Also

- ▶ Diffusion of Technology
- ▶ General Purpose Technologies
- ▶ Information Technology and the World Economy
- ▶ Internet, Economics of the
- ▶ Technical Change

Bibliography

Bresnahan, T., and S. Greenstein. 1999. Technological competition and the structure of the computer industry. *Journal of Industrial Economics* 47(1): 1–40.

Brensnahan, T., and F. Malerba. 1999. Industrial dynamics and the evolution of firm's and nations' competitive capabilities in the world computer industry. In *Sources of industrial leadership*, ed. D. Mowery and R. Nelson. Cambridge, UK: Cambridge University Press.

Brynjolfsson, E., and H. Lorin. 2000. Beyond computation: Information technology, organizational transformation and business performance. *Journal of Economic Perspectives* 14(4): 23–48.

Dedrick, J., and K. Kraemer. 2005. The impacts of IT on firm and industry structure: The personal computer industry. *California Management Review* 47(3): 122–142.

Flamm, K. 2003. The new economy in historical perspective: Evolution of digital electronics technology. In *New economy handbook*, ed. D. Jones. San Diego: Academic Press/Elsevier.

Greenstein, S. (ed.). 2006. *The industrial economics of computing*. Northampton: Edward Elgar.

Jorgenson, D., and C. Wessner (eds.). 2005. *Deconstructing the computer: Report of a symposium*. Washington, DC: National Academies Press.

McKinsey Global Institute. 2001. *U.S. productivity growth, 1995–2000: Understanding the contribution of information technology relative to other factors*. Washington, DC: McKinsey and Co.

Computer Science and Game Theory

Joseph Y. Halpern

Abstract

Work at the intersection of computer science and game theory is briefly surveyed, with a focus on the work in computer science. In particular, the following topics are considered: various roles of computational complexity in game theory, including modelling bounded rationality, its role in mechanism design, and the problem of computing Nash equilibria; the *price of anarchy,* that is, the cost of using decentralizing solution to a problem; and interactions between distributed computing and game theory.

Keywords

Algorithmic knowledge; Algorithmic mechanism design; Bayesian networks; Bounded rationality; Byzantine agreement; Cheap talk;

Coalitions; Combinatorial auctions; Complexity theory; Computational complexity; Computer science and game theory; Distributed computing; Efficient representation of games; Game theory; Gibbard–Satterthwaite th; Implementing mediators; Interactive epistemology; k-resilient equilibrium; (k,t)-robust equilibrium; Learning; Mechanism design; Markov networks; Nash equilibrium; Price of anarchy; Prisoner's Dilemma; Regret; Strategic voting; Tit for tat; Voting

JEL Classifications
C0; C6

Introduction

There has been a remarkable increase in work at the interface of computer science and game theory in the past decade. Game theory forms a significant component of some major computer science conferences (see, for example, Kearns and Reiter 2005; Sandholm and Yokoo 2003); leading computer scientists are often invited to speak at major game theory conferences, such as the World Congress on Game Theory 2000 and 2004. In this article I survey some of the main themes of work in the area, with a focus on the work in computer science. Given the length constraints, make no attempt at being comprehensive, especially since other surveys are also available, including Halpern (2003), Linial (1994), Papadimitriou (2001), and a comprehensive survey book (Nisan et al. 2007).

The survey is organized as follows. I look at the various roles of computational complexity in game theory in section "Complexity Considerations", including its use in modelling bounded rationality, its role in mechanism design, and the problem of computing Nash equilibria. In section "The Price of Anarchy", I consider a game-theoretic problem that originated in the computer science literature, but should be of interest to the game theory community: computing the *price of anarchy,* that is, the cost of using a decentralizing solution to a problem. In section "Game Theory and Distributed Computing", I consider interactions between distributed computing and game theory. In section "Implementing Mediators", I consider the problem of implementing mediators, which has been studied extensively in both computer science and game theory. I conclude in section "Other Topics" with a discussion of a few other topics of interest.

Complexity Considerations

The influence of computer science in game theory has perhaps been most strongly felt through complexity theory. I consider some of the strands of this research here. There are a numerous basic texts on complexity theory that the reader can consult for more background on notions like NP-completeness and finite automata, including Hopcroft and Ullman (1979) and Papadimitriou (1994a).

Bounded Rationality

One way of capturing bounded rationality is in terms of agents who have limited computational power. In economics, this line of research goes back to the work of Neyman (1985) and Rubinstein (1986), who focused on finitely repeated Prisoner's Dilemma. In n-round finitely repeated Prisoner's Dilemma, there are 2^{2n-1} strategies (since a strategy is a function from histories to {cooperate, defect}, and there are clearly $2^n - 1$ histories of length $< n$). Finding a best response to a particular move can thus potentially be difficult. Clearly people do not find best responses by doing extensive computation. Rather, they typically rely on simple heuristics, such as 'tit for tat' (Axelrod 1984). Such heuristics can often be captured by finite automata; both Neyman and Rubinstein thus focus on finite automata playing repeated Prisoner's Dilemma. Two computer scientists, Papadimitriou and Yannakakis (1994), showed that if both players in an n-round Prisoner's Dilemma are finite automata with at least $2^n - 1$ states, then the only equilibrium is the one where they defect in every round. This result says that a finite automaton with exponentially many states

can compute best responses in Prisoner's Dilemma.

We can then model bounded rationality by restricting the number of states of the automaton. Neyman (1985) showed, roughly speaking, that if the two players in *n*-round Prisoner's Dilemma are modelled by finite automata with a number of states in the interval $[n^{1/k}, n^k]$ for some *k,* then collaboration can be approximated in equilibrium; more precisely, if the payoff for (cooperate, cooperate) is (3, 3) there is an equilibrium in the repeated game where the average payoff per round is greater than $3 - \frac{1}{k}$ for each player. Papadimitriou and Yannakakis (1994) sharpen this result by showing that if at least one of the players has fewer than $2^{c_\varepsilon n}$ states, where $c_\varepsilon = \frac{\varepsilon}{12(1+\varepsilon)}$, then for sufficiently large n, there is an equilibrium where each player's average payoff per round is greater than $3 - \varepsilon$. Thus, computational limitations can lead to cooperation in Prisoner's Dilemma.

There have been a number of other attempts to use complexity-theoretic ideas from computer science to model bounded rationality (see Rubinstein 1998, for some exs). However, it seems that there is much more work to be done here.

Computing Nash Equilibrium

Nash (1950) showed every finite game has a Nash equilibrium in mixed strategies. But how hard is it to actually find that equilibrium? On the positive side, there are well known algorithms for computing Nash equilibrium, going back to the classic Lemke–Howson (1964) algorithm, with a spate of recent improvements (see, for example, Govindan and Wilson 2003; Blum et al. 2003; Porter et al. 2004). Moreover, for certain classes of games (for example, symmetric games, Papadimitriou and Roughgarden 2005), there are known to be polynomial-time algorithms. On the negative side, many qsts about Nash equilibrium are known to be NP-hard. For example, Gilboa and Zemel (1989) showed that, for a game presented in normal form, deciding whether there exists a Nash equilibrium where each player gets a payoff of at least *r* is NP-complete. Interestingly, Gilboa and Zemel also show that computing whether there exists a *correlated* equilibrium (Aumann 1987) where each player gets a payoff of at least *r* is computable in polynomial time. In general, qsts regarding correlated equilibrium seem easier than the analogous qsts for Nash equilibrium; see Papadimitriou (2005) and Papadimitriou and Roughgarden (2005) for further examples. Chu and Halpern (2001) prove similar NP-completeness results if the game is represented in extensive form, even if all players have the same payoffs (a situation that arises frequently in computer science applications, where we can view the players as agents of some designer, and take the payoffs to be the designer's payoffs). Conitzer and Sandholm (2003) give a compendium of hardness results for various qsts regarding Nash equilibria.

Nevertheless, there is a sense in which it seems that the problem of finding a Nash equilibrium is easier than typical NP-complete problems, because every game is guaranteed to have a Nash equilibrium. By way of contrast, for a typical NP- complete problem like prptal satisfiability, whether or not a prptal formula is satisfiable is not known. Using this observation, it can be shown that if finding a Nash equilibrium is NP-complete, then NP = coNP. Recent work has in a sense completely characterized the complexity of finding a Nash equilibrium in normal-form games: it is a *PPAD-complete* problem (Chen and Deng 2006; Daskalis et al. 2006). PPAD stands for 'polynomial parity argument (directed case)'; see Papadimitriou (1994b) for a formal definition and examples of other PPAD problems. It is believed that PPAD-complete problems are not solvable in polynomial time, but are simpler than NP-complete problems, although this remains an open problem. See Papadimitriou (2007) for an overview of this work.

Algorithmic Mechanism Design

The problem of mechanism design is to design a game such that the agents playing the game, motivated only by self-interest, achieve the designer's goals. This problem has much in common with the standard computer science problem of designing protocols that satisfy certain specifications (for example, designing a distributed protocol that

achieves Byzantine agreement; see section "Game Theory and Distributed Computing"). Work on mechanism design has traditionally ignored computational concerns. But Kfir-Dahav et al. (2000) show that, even in simple settings, optimizing social welfare is NP-hard, so that perhaps the most common approach to designing mechanisms, applying the Vickrey–Groves–Clarke (VCG) procedure (Clarke 1971; Groves 1973; Vickrey 1961), is not going to work in large systems. We might hope that, even if we cannot compute an optimal mechanism, we might be able to compute a reasonable approximation to it. However, as Nisan and Ronen (2000, 2001) show, in general, replacing a VCG mechanism by an approximation does not preserve truthfulness. That is, even though truthfully revealing one's type is an optimal strategy in a VCG mechanism, it may no longer be optimal in an approximation. Following Nisan and Ronen's work, there has been a spate of papers either describing computationally tractable mechanisms or showing that no computationally tractable mechanism exists for a number of problems, ranging from task allocation (Archer and Tardos 2001; Nisan and Ronen 2001) to cost-sharing for multicast trees (Feigenbaum et al. 2000) (where the problem is to share the cost of sending, for example, a movie over a network among the agents who actually want the movie) to finding low-cost paths between nodes in a network (Archer and Tardos 2002).

The problem that has attracted perhaps the most attention is *combinatorial auctions*, where bidders can bid on bundles of items. This becomes of particular interest in situations where the value to a bidder of a bundle of goods cannot be determined by simply summing the value of each good in isolation. To take a simple example, the value of a pair of shoes is much higher than that of the individual shoes; perhaps more interestingly, an owner of radio stations may value having a licence in two adjacent cities more than the sum of the individual licences. Combinatorial auctions are of great interest in a variety of settings including spectrum auctions, airport time slots (that is, take-off and landing slots), and industrial procurement. There are many complexity-theoretic issues related to combinatorial auctions. For a detailed discussion and references see Cramton et al. (2006); I briefly discuss a few of the issues involved here.

Suppose that there are n items being auctioned. Simply for a bidder to communicate her bids to the auctioneer can take, in general, exponential time, since there are 2^n bundles. In many cases, we can identify a bid on a bundle with the bidder's valuation of the bundle. Thus, we can try to carefully design a bidding language in which a bidder can communicate her valuations succinctly. Simple information-theoretic arguments can be used to show that, for every bidding language, there will be valuations that will require length at least 2^n to express in that language. Thus, the best we can hope for is to design a language that can represent the 'interesting' bids succinctly. See Nisan (2006) for an overview of various bidding languages and their expressive power.

Given bids from each of the bidders in a combinatorial auction, the auctioneer would like to then determine the winners. More precisely, the auctioneer would like to allocate the m items in an auction so as to maximize his revenue. This problem, called the *winner determination problem*, is NP-complete in general, even in relatively simple classes of combinatorial auctions with only two bidders making rather restricted bids. Moreover, it is not even polynomial-time approximable, in the sense that there is no constant d and polynomial-time algorithm such that the algorithm produces an allocation that gives revenue that is at least 1/d of optimal. On the other hand, there are algorithms that provably find a good solution, seem to work well in practice, and, if they seem to be taking too long, can be terminated early, usually with a good feasible solution in hand. See Lehmann et al. (2006), for an overview of the results in this area.

In most mechanism design problems, computational complexity is seen as the enemy. There is one class of problems in which it may be a friend: voting. One problem with voting mechanisms is that of *manipulation* by voters. That is, voters may be tempted to vote strategically rather than ranking the candidates according to their true preferences, in the hope that the final outcome will be

more favourable. This situation arises frequently in practice; in the 2000 US presidential election, American voters who preferred Nader to Gore to Bush were encouraged to vote for Gore, rather than 'wasting' a vote on Nader. The classic Gibbard–Satterthwaite theorem (Gibbard 1973; Satterthwaite 1975) shows that, if there are at least three alternatives, then in any nondictatorial voting scheme (that is, one where it is *not* the case that one particular voter dictates the final outcome, irrespective of how the others vote), there are preferences under which an agent is better off voting strategically. The hope is that, by constructing the voting mechanism appropriately, it may be computationally intractable to find a manipulation that will be beneficial. While finding manipulations for the plurality protocol (the candidate with the most votes wins) is easy, there are well-known voting protocols for which manipulation is hard in the presence of three or more candidates. See Conitzer et al. (2007) for a summary of results and further pointers to the literature.

Communication Complexity

Most mechanisms in the economics literature are designed so that agents truthfully reveal their preferences. However, in some settings, revealing one's full preferences can require a prohibitive amount of communication. For example, in a combinatorial auction of m items, revealing one's full preferences may require revealing what one would be willing to pay for each of the $2^m - 1$ possible bundles of items. Even if $m = 30$, this requires revealing more than one billion numbers. This leads to an obvious qst: how much communication is required by various mechanisms? Formal work on this question in the economics community goes back to Hurwicz (1977) and Mount and Reiter (1974); their definitions focused on the dimension of the message space. Independently (and later), there was active work in computer science on *communication complexity*, the number of bits of communication needed for a set of n agents to compute the value of a function $f : x_{i=1}^{n}\Theta_i \rightarrow X$, where each agent i knows $\theta_i \in \Theta_i$ (Think of θ_i as representing agent i's type.) Recently there has been an explosion of work, leading to a better understanding of the communication complexity for many important economic allocation problems; see Segal (2006) for an overview. Two important themes in this work are understanding the role of price- based market mechanisms in solving allocation problems with minimal communication, and designing mechanisms that provide agents with incentives to communicate truthfully while having low communication requirements.

The Price of Anarchy

In a computer system, there are situations where we may have a choice between a centralized and a decentralized solution to a problem. By 'centralized' here, I mean that each agent in the system is told exactly what to do and must do so; in the decentralized solution, each agent tries to optimize his own selfish interests. Of course, centralization comes at a cost. For one thing, there is a problem of enforcement. For another, centralized solutions tend to be more vulnerable to failure. On the other hand, a centralized solution may be more socially beneficial. How much more beneficial can it be?

Koutsoupias and Papadimitriou (1999) formalized this question by considering the ratio of the social welfare of the centralized solution to the social welfare of the Nash equilibrium with the worst social welfare (assuming that the social welfare function is always positive). They called this ratio the *price of anarchy,* and proved a number of results regarding the price of anarchy for a scheduling problem on parallel machines. Since the original paper, the price of anarchy has been studied in many settings, including traffic routing (Roughgarden and Tardos 2002), facility location games (for example, where is the best place to put a factory) (Vetta 2002), and spectrum sharing (how should channels in a WiFi network be assigned) (Halldórsson et al. 2004).

To give a sense of the results, consider the traffic-routing context of Roughgarden and Tardos (2002). Suppose that the travel time on a road increases in a known way with the congestion on the road. The goal is to minimize the

average travel time for all drivers. Given a road network and a given traffic load, a centralized solution would tell each driver which road to take. For example, there could be a rule that cars with odd-numbered licence plates take road 1, while those with even-numbered plates take road 2, to minimize congestion on either road. Roughgarden and Tardos show that the price of anarchy is unbounded if the travel time can be a nonlinear function of the congestion. On the other hand, if it is linear, they show that the price of anarchy is at most 4/3.

The price of anarchy is but one way of computing the 'cost' of using a Nash equilibrium. Others have been considered in the computer science literature. For example, Tennenholtz (2002) compares the *safety level* of a game – the optimal amount that an agent can guarantee himself, independent of what the other agents do – to what the agent gets in a Nash equilibrium, and shows, for interesting classes of games, including load-balancing games and first-price auctions, that the ratio between the safety level and the Nash equilibrium is bounded. For example, in the case of first- price auctions, it is bounded by the constant e.

Game Theory and Distributed Computing

Distributed computing and game theory are interested in much the same problems: dealing with systems where there are many agents, facing uncertainty and having possibly different goals. In practice, however, there has been a significant difference in emphasis between the two areas. In distributed computing, the focus has been on problems such as fault tolerance, asynchrony, scalability, and proving correctness of algorithms; in game theory, the focus has been on strategic concerns. I discuss here some issues of common interest. Most of the discussion in the remainder of this section is taken from Halpern (2003).

To understand the relevance of fault tolerance and asynchrony, consider the *Byzantine agreement* problem, a paradigmatic problem in the distributed systems literature. In this problem, there are assumed to be n soldiers, up to t of which may be faulty (the t stands for *traitor*); n and t are assumed to be common knowledge. Each soldier starts with an initial preference, to either attack or retreat. (More precisely, there are two types of nonfaulty agents – those that prefer to attack, and those that prefer to retreat.) We want a protocol that guarantees that (1) all *nonfaulty* soldiers reach the same decision, and (2) if all the soldiers are nonfaulty and their initial preferences are identical, then the final decision agrees with their initial preferences. (The condition simply prevents the obvious trivial solutions, where the soldiers attack no matter what, or retreat no matter what.)

The problem was introduced by Pease et al. (1980), and has been studied in detail since then; Chor and Dwork (1989), Fischer (1983), and Linial (1994) provide overviews. Whether the Byzantine agreement problem is solvable depends in part on what types of failures are considered, on whether the system is *synchronous* or *asynchronous*, and on the ratio of n to t. Roughly speaking, a system is synchronous if there is a global clock and agents move in lockstep; a 'step' in the system corresponds to a tick of the clock. In an asynchronous system, there is no global clock. The agents in the system can run at arbitrary rates relative to each other. One step for agent 1 can correspond to an arbitrary number of steps for agent 2 and vice versa. Synchrony is an implicit assumption in essentially all games. Although it is certainly possible to model games where player 2 has no idea how many moves player 1 has taken when player 2 is called upon to move, it is not typical to focus on the effects of synchrony (and its lack) in games. On the other hand, in distributed systems, it is typically a major focus.

Suppose for now that we restrict to *crash failures*, where a faulty agent behaves according to the protocol, except that it might crash at some point, after which it sends no messages. In the round in which an agent fails, the agent may send only a subset of the messages that it is supposed to send according to its protocol. Further suppose that the system is synchronous. In this case, the following rather simple protocol achieves Byzantine agreement:

- In the first round, each agent tells every other agent its initial preference.
- For rounds 2 to $t + 1$, each agent tells every other agent everything it has heard in the previous round. Thus, for example, in round 3, agent 1 may tell agent 2 that it heard from agent 3 that its initial preference was to attack, and that it (agent 3) heard from agent 2 that its initial preference was to attack, and it heard from agent 4 that its initial preferences was to retreat, and so on. This means that messages get exponentially long, but it is not difficult to represent this information in a compact way so that the total communication is polynomial in n, the number of agents.
- At the end of round $t + 1$, if an agent has heard from any other agent (including itself) that its initial preference was to attack, it decides to attack; otherwise, it decides to retreat.

Why is this correct? Clearly, if all agents are correct and want to retreat (resp., attack), then the final decision will be to retreat (resp., attack), since that is the only preference that agents hear about (recall that for now we are considering only crash failures). It remains to show that if some agents prefer to attack and others to retreat, then all the nonfaulty agents reach the same final decision. So suppose that i and j are nonfaulty and i decides to attack. That means that i heard that some agent's initial preference was to attack. If it heard this first at some round $t' < t + 1$, then i will forward this message to j, who will receive it and thus also attack. On the other hand, suppose that i heard it first at round $t + 1$ in a message from i_{t+1}. Thus, this message must be of the form *'*i_t said at round t that... that i_2 said at round 2 that i_1 said at round 1 that its initial preference was to attack.' Moreover, the agents $i_1, \ldots, i_{t+1}$ must all be distinct. Indeed, it is easy to see that i_k must crash in round k before sending its message to i (but after sending its message to i_{k+1}), for $k = 1, \ldots, t$, for otherwise i must have gotten the message from i_k, contradicting the assumption that i first heard at round $t + 1$ that some agent's initial preference was to attack. Since at most t agents can crash, it follows that i_{t+1}, the agent that sent the message to i, is not faulty, and thus sends the message to j. Thus, j also decides to attack. A symmetric argument shows that if j decides to attack, then so does i.

It should be clear that the correctness of this protocol depends on both the assumptions made: crash failures and synchrony. Suppose instead that *Byzantine* failures are allowed, so that faulty agents can deviate in arbitrary ways from the protocol; they may 'lie', send deceiving messages, and collude to fool the nonfaulty agents in the most malicious ways. In this case, the protocol will not work at all. In fact, it is known that agreement can be reached in the presence of Byzantine failures iff $t < n/3$, that is, iff fewer than a third of the agents can be faulty (Pease et al. 1980). The effect of asynchrony is even more devastating: in an asynchronous system, it is impossible to reach agreement using a deterministic protocol even if $t = 1$ (so that there is at most one failure) and only crash failures are allowed (Fischer et al. 1985). The problem in the asynchronous setting is that if none of the agents have heard from, say, agent 1, they have no way of knowing whether agent 1 is faulty or just slow. Interestingly, there are randomized algorithms (that is, behavioural strategies) that achieve agreement with arbitrarily high probability in an asynchronous setting (Ben-Or 1983; Rabin 1983).

Byzantine agreement can be viewed as a game where, at each step, an agent can either send a message or decide to attack or retreat. It is essentially a game between two teams, the nonfaulty agents and the faulty agents, whose composition is unknown (at least by the correct agents). To model it as a game in the more traditional sense, we could imagine that the nonfaulty agents are playing against a new player, the 'adversary'. One of the adversary's moves is that of 'corrupting' an agent: changing its type from 'nonfaulty' to 'faulty.' Once an agent is corrupted, what the adversary can do depends on the failure type being considered. In the case of crash failures, the adversary can decide which of a corrupted agent's messages will be delivered in the round in which the agent is corrupted; however, it cannot modify the messages themselves. In the case of Byzantine failures, the adversary essentially gets to make the moves for agents that have been

corrupted; in particular, it can send arbitrary messages.

Why has the distributed systems literature not considered strategic behaviour in this game? Crash failures are used to model hardware and software failures; Byzantine failures are used to model random behaviour on the part of a system (for example, messages getting garbled in transit), software errors, and malicious adversaries (for example, hackers). With crash failures, it does not make sense to view the adversary's behaviour as strategic, since the adversary is not really viewed as having strategic interests. While it would certainly make sense, at least in principle, to consider the probability of failure (that is, the probability that the adversary corrupts an agent), this approach has by and large been avoided in the literature because it has proved difficult to characterize the probability distribution of failures over time. Computer components can perhaps be characterized as failing according to an exponential distribution (see Babaoglu 1987, for an analysis of Byzantine agreement in such a setting), but crash failures can be caused by things other than component failures (faulty software, for ex); these can be extremely difficult to characterize probabilistically. The problems are even worse when it comes to modelling random Byzantine behaviour.

With malicious Byzantine behaviour, it may well be reasonable to impute strategic behaviour to agents (or to an adversary controlling them). However, it is often difficult to characterize the payoffs of a malicious agent. The goals of the agents may vary from that of simply trying to delay a decision to that of causing disagreement. It is not clear what the appropriate payoffs should be for attaining these goals. Thus, the distributed systems literature has chosen to focus instead on algorithms that are guaranteed to satisfy the specification without making assumptions about the adversary's payoffs (or nature's probabilities, in the case of crash failures).

Recently, there has been some work on adding strategic concerns to standard problems in distributed computing; see, for example, Alvisi et al. (2005) and Halpern and Teague (2004). Moving in the other direction, there has also been some work on adding concerns of fault tolerance and asynchrony to standard problems in game theory; see, for example, Eliaz (2002), Monderer and Tennenholtz (1999a, b) and the definitions in the next section. This seems to be an area that is ripe for further developments. One such development is the subject of the next section.

C

Implementing Mediators

The question of whether a problem in a multiagent system that can be solved with a trusted mediator can be solved by just the agents in the system, without the mediator, has attracted a great deal of attention in both computer science (particularly in the cryptography community) and game theory. In cryptography, the focus on the problem has been on *secure multiparty computation.* Here it is assumed that each agent i has some private information x_i. Fix functions $f_1, \ldots, f_n$. The goal is to have agent i learn $f_i(x_1, \ldots, x_n)$ without learning anything about Xj for $j \neq i$ beyond what is revealed by the value off $(x_1 \ldots, x_n)$. With a trusted mediator, this is trivial: each agent i just gives the mediator its private value x_i the mediator then sends each agent i the value $f_i(x_1, \ldots, x_n)$. Work on multiparty computation (Goldreich et al. 1987; Shamir et al. 1981; Yao 1982) provides conditions under which this can be done. In game theory, the focus has been on whether an equilibrium in a game with a mediator can be implemented using what is called *cheap talk* – that is, just by players communicating among themselves (cf. Barany 1992; Ben-Porath 2003; Forges 1990; Gerardi 2004; Heller 2005; Urbano and Vila 2004). As suggested in the previous section, the focus in the computer science literature has been in doing multiparty computation in the presence of possibly malicious adversaries, who do everything they can to subvert the computation, while in the game theory literature the focus has been on strategic agents. In recent work, Abraham et al. (2006) and Abraham et al. (2007) considered deviations by both rational players, who have preferences and try to maximize them, and players who can viewed as malicious, although it is perhaps better to think

of them as rational players whose utilities are not known by the other players or mechanism designer. I briefly sketch their results here; the following discussion is taken from Abraham et al. (2007).

The idea of tolerating deviations by coalitions of players goes back to Aumann (1959); more recent refinements have been considered by Moreno and Wooders (1996). Aumann's definition is essentially the following.

Definition 1 *$\sigma^{\Rightarrow}$ is a k-resilient' equilibrium if, for all sets C of players with $|C| \leq k$, it is not the case that there exists a strategy $\tau^{\Rightarrow}$ such that $u_i(\tau^{\Rightarrow}{}_C, \sigma^{\Rightarrow}{}_{-C}) > u_i(\sigma^{\Rightarrow})$ for all $i \in$ C.*

As usual, the strategy $(\tau^{\Rightarrow}{}_C, \sigma^{\Rightarrow}{}_{-C})$ is the one where each player $i \in$ C plays τ_i and each player $i \notin C$ plays σ_i. As the prime notation suggests, this is not quite the definition we want to work with. The trouble with this definition is that it suggests that coalition members cannot communicate with each other during the game. Perhaps surprisingly, allowing communication can *prevent* certain equilibria (see Abraham et al. 2007, for an ex). Since we should expect coalition members to communicate, the following definition seems to capture a more reasonable notion of resilient equilibrium. Let the cheap-talk extension of a game Γ be, roughly speaking, the game where players are allowed to communicate among themselves in addition to performing the actions of Γ and the payoffs are just as in Γ.

Definition 2 $\sigma^{\Rightarrow}$ is a *k-resilient equilibrium* in a game Γ if $\sigma^{\Rightarrow}$ is a k-resilient' equilibrium in the cheap-talk extension of Γ (where we identify the strategy σ_t in the game Γ with the strategy in the cheap-talk game where player i never sends any messages beyond those sent according to σ_i).

A standard assumption in game theory is that utilities are (commonly) known; when we are given a game we are also given each player's utility. When players make decisions, they can take other players' utilities into account. However, in large systems it seems almost invariably the case that there will be some fraction of users who do not respond to incentives the way we expect. For example, in a peer-to-peer network like Kazaa or Gnutella, it would seem that no rational agent should share files. Whether or not you can get a file depends only on whether other people share files. Moreover, there are disincentives for sharing (the possibility of lawsuits, use of bandwidth, and so on). Nevertheless, people do share files. However, studies of the Gnutella network have shown almost 70 per cent of users share no files and nearly 50 per cent of responses are from the top one per cent of sharing hosts (Adar and Huberman 2000).

One reason that people might not respond as we expect is that they have utilities that are different from those we expect. Alternatively, the players may be irrational, or (if moves are made using a computer) they may be playing using a faulty computer and thus not able to make the move they would like, or they may not understand how to get the computer to make the move they would like. Whatever the reason, it seems important to design strategies that tolerate such unanticipated behaviours, so that the payoffs of the users with 'standard' utilities do not get affected by the nonstandard players using different strategies. This can be viewed as a way of adding fault tolerance to equilibrium notions.

Definition 3 A joint strategy $\sigma^{\Rightarrow}$ is *t-immune* if, for all T $\subseteq$ N with |T| $\leq$ t, all joint strategies $\tau^{\Rightarrow}$, and all $i \notin T$, we have $u_i(\sigma^{\Rightarrow}{}_{-T}, \tau^{\Rightarrow}{}_T) \geq u_i(\sigma^{\Rightarrow})$.

The notion of *t*-immunity and *k*-resilience address different concerns. For *t* immunity, we consider the payoffs of the players not in T, and require that they are not worse due to deviation; for resilience, we consider the payoffs of players in *C*, and require that they are not better due to deviation. It is natural to combine both notions. Given a game Γ, let Γ.be the game that is identical to Γ except that the players in *T* are fixed to playing strategy τ.

Definition 4 $\sigma^{\Rightarrow}$ is a (*k*, *t*) *-robust* equilibrium if $\sigma^{\Rightarrow}$ is t-immune and, for all T $\subseteq$ N such that $|T| \leq t$ and all joint strategies $\tau^{\Rightarrow}$, $\sigma^{\Rightarrow}{}_{-T}$ is a *k*-resilient strategy of $\Gamma_T^{\tau^{\Rightarrow}}$.

To state the results of Abraham et al. (2006) and (2007) on implementing mediators, three games need to be considered: an *underlying*

game Γ, an extension Γ_d of Γ with a mediator, and a cheap-talk extension Γ_{CT} of Γ. Assume that Γ is a *normal-form Bayesian game:* each player has a type from some type space with a known distribution over types, and the utilities of the agents depend on the types and actions taken. Roughly speaking, a cheap talk game *implements* a game with a mediator if it induces the same distribution over actions in the underlying game, for each type vector of the players. With this background, I can summarize the results of Abraham et al. (2006) and (2007).

- If $n > 3k + 3t$, a (k, t) -robust strategy $\sigma^{\Rightarrow}$ with a mediator can be implemented using cheap talk (that is, there is a (k, t)-robust strategy $\sigma^{\Rightarrow\prime}$ in a cheap talk game such that $\sigma^{\Rightarrow}$ and $\sigma^{\Rightarrow\prime}$ induce the same distribution over actions in the underlying game). Moreover, the implementation requires no knowledge of other agents' utilities, and the cheap talk protocol has bounded running time that does not depend on the utilities.
- If $n \leq 3k + 3t$, then, in general, mediators cannot be implemented using cheap talk without knowledge of other agents' utilities. Moreover, even if other agents' utilities are known, mediators cannot, in general, be implemented without having a $(k + t)$-punishment strategy (that is, a strategy that, if used by all but at most $(k + t)$ players, guarantees that every player gets a worse outcome than they do with the equilibrium strategy) nor with bounded running time.
- If $n > 2k + 3t$, then mediators can be implemented using cheap talk if there is a punishment strategy (and utilities are known) in finite expected running time that does not depend on the utilities.
- If $n \# 2k + 3t$ then mediators cannot, in general, be implemented, even if there is a punishment strategy and utilities are known.
- If $n > 2k + 2t$ and there are broadcast channels then, for all ε, mediators can be ε -implemented (intuitively, there is an implementation where players get utility within e of what they could get by deviating) using cheap talk, with bounded expected running time that does not depend on the utilities.
- If $n \leq 2k + 2t$, then mediators cannot, in general, be e-implemented, even with broadcast channels. Moreover, even assuming cryptography and polynomially bounded players, the expected running time of an implementation depends on the utility functions of the players and ε.
- If $n > k + 3$t, then, assuming cryptography and polynomially bounded players, mediators can be e-implemented using cheap talk, but if $n \# 2k + 2t$, then the running time depends on the utilities in the game and ε.
- If $n \leq k + 3t$, then even assuming cryptography, polynomially bounded players, and a $(k + t)$ -punishment strategy, mediators cannot, in general, be ε -implemented using cheap talk.
- If $n > k + t$, then, assuming cryptography, polynomially bounded players, and a public-key infrastructure (PKI), we can ε-implement a mediator.

The proof of these results makes heavy use of techniques from computer science. All the possibility results showing that mediators can be implemented use techniques from secure multiparty computation. The results showing that if $n \leq 3k + 3$t, then we cannot implement a mediator without knowing utilities, and that, even if utilities are known, a punishment strategy is required, use the fact that Byzantine agreement cannot be reached if $t < n/3$; the impossibility result for $n \leq 2k + 3t$ also uses a variant of Byzantine agreement.

A related line of work considers implementing mediators assuming stronger primitives (which cannot be implemented in computer networks); see Izmalkov et al. (2005) and Lepinski et al. (2004) for details.

Other Topics

There are many more areas of interaction between computer science than I have indicated in this brief survey. I briefly mention a few others here.

Interactive Epistemology

Since the publication of Aumann's (1976) seminal paper, there has been a great deal of activity in

trying to understand the role of knowledge in games, and providing epistemic analyses of solution concepts; see Battigalli and Bonanno (1999) for a survey. In computer science, there has been a parallel literature applying epistemic logic to reason about distributed computation. One focus of this work has been on characterizing the level of knowledge needed to solve certain problems. For example, to achieve Byzantine agreement common knowledge among the nonfaulty agents of an initial value is necessary and sufficient. More generally, in a precise sense, common knowledge is necessary and sufficient for coordination. Another focus has been on defining logics that capture the reasoning of resource-bounded agents. A number of approaches have been considered. Perhaps the most common considers logics for reasoning about *awareness*, where an agent may not be aware of certain concepts, and can know something only if he is aware of it. This topic has been explored in both computer science and game theory; see Dekel et al. (1998), Fagin and Halpern (1988), Halpern (2001), Halpern and Rêgo (2007), Heifetz et al. (2006), and Modica and Rustichini (1994, 1999) for some of the work in this active area. Another approach, so far considered only by computer scientists, involves *algorithmic knowledge*, which takes seriously the assumption that agents must explicitly compute what they know. See Fagin et al. (1995) for an overview of the work in epistemic logic in computer science.

Network Growth

If we view networks as being built by selfish players (who decide whether or not to build links), what will the resulting network look like? How does the growth of the network affect its functionality? For example, how easily will influence spread through the network? How easy is it to route traffic? See Fabrikant et al. (2003) and Kempe et al. (2003) for some recent computer science work in this burgeoning area.

Efficient Representation of Games

Game theory has typically focused on 'small' games, often two- or three-player games, that are easy to describe, such as Prisoner's Dilemma, in order to understand subtleties regarding basic issues such as rationality. To the extent that game theory is used to tackle larger, more practical problems, it will become important to find efficient techniques for describing and analysing games. By way of analogy, $2^n - 1$ numbers are needed to describe a probability distribution on a space characterized by n binary random variables. For $n = 100$ (not an unreasonable number in practical situations), it is impossible to write down the probability distribution in the obvious way, let alone do computations with it. The same issues will surely arise in large games. Computer scientists use graphical approaches, such as *Bayesian networks* and *Markov networks* (Pearl 1988), for representing and manipulating probability measures on large spaces. Similar techniques seem applicable to games; see, for example, Kearns et al. (2001), Koller and Milch (2001), and La Mura (2000) for specific approaches, and Kearns (2007) for a recent overview. Note that representation is also an issue when we consider the complexity of problems such as computing Nash or correlated equilibria. The complexity of a problem is a function of the size of the input, and the size of the input (which in this case is a description of the game) depends on how the input is represented.

Learning in Games

There has been a great deal of work in both computer science and game theory on learning to play well in different settings (see Fudenberg and Levine 1998, for an overview of the work in game theory). One line of research in computer science has involved learning to play optimally in a reinforcement-learning setting, where an agent interacts with an unknown (but fixed) environment. The agent then faces a fundamental tradeoff between *exploration* and *exploitation.* The question is how long it takes to learn to play well (to get a reward within some fixed ε of optimal); see Brafman and Tennenholtz (2002) and Kearns and Singh (1998) for the current state of the art. A related question is efficiently finding a strategy minimizes *regret* – that is, finding a strategy that is guaranteed to do not much worse than the best strategy would have done in hindsight (that is,

even knowing what the opponent would have done). See Blum and Mansour (2007) for a recent overview of work on this problem.

See Also

- ▶ Computation of General Equilibria
- ▶ Computational Methods in Econometrics
- ▶ Computing in Mechanism Design
- ▶ Data Mining
- ▶ Electronic Commerce
- ▶ Epistemic Game Theory: An Overview
- ▶ Epistemic Game Theory: Beliefs and Types
- ▶ Mathematics of Networks
- ▶ Mechanism Design (New Developments)
- ▶ Rationality, Bounded
- ▶ Voting Paradoxes

Acknowledgment *The work for this article was supported in part by NSF under grants CTC-0208535 and ITR-0325453, by ONR under grant N00014-02-1-0455, by the DoD Multidisciplinary University Research Initiative (MURI) program administered by the ONR under grants N00014-01-1-0795 and N00014-04-1-0725, and by AFOSR under grant F49620-02-1-0101. Thanks to Larry Blume, Christos Papadimitriou, Ilya Segal, Éva Tardos, and Moshe Tennenholtz for useful comments*.

Bibliography

Abraham, I., D. Dolev, R. Gonen, and J. Halpern. 2006. Distributed computing meets game theory: Robust mechanisms for rational secret sharing and multiparty computation. In *Proceedings of the 25th ACM symposium on principles of distributed computing*, 53–62.

Abraham, I., Dolev, D., and Halpern, J. 2007. *Lower bounds on implementing robust and resilient mediators*. Available at http://arxiv.org/abs/0704.3646. Accessed 19 June 2007.

Adar, E., and B. Huberman. 2000. Free riding on Gnutella. *First Monday* 5(10).

Alvisi, L., A.S. Ayer, A. Clement, M. Dahlin, J.P. Martin, and C. Porth. 2005. BAR fault tolerance for cooperative services. In *Proceedings of the 20th ACM symposium on operating systems principles* (*SOSP 2005*), 45–58.

Archer, A., and É. Tardos. 2001. Truthful mechanisms for one-parameter agents. In *Proceedings of the 42nd IEEE symposium on foundations of computer science*, 482–491.

Archer, A., and É. Tardos. 2002. Frugal path mechanisms. In *Proceedings of the 13th ACM-SIAM symposium on discrete algorithms*, 991–999.

Aumann, R. 1959. Acceptable points in general cooperative n-person games. In *Contributions to the theory of games*, ed. A. Tucker and R. Luce, vol. IV, 287–324. Princeton: Princeton University Press.

Aumann, R.J. 1976. Agreeing to disagree. *Annals of Statistics* 4: 1236–1239.

Aumann, R.J. 1987. Correlated equilibrium as an expression of Bayesian rationality. *Econometrica* 55: 1–18.

Axelrod, R. 1984. *The evolution of cooperation*. New York: Basic Books.

Babaoglu, O. 1987. On the reliability of consensus-based fault-tolerant distributed computing systems. *ACM Translation on Computer Systems* 5: 394–416.

Barany, I. 1992. Fair distribution protocols or how the players replace fortune. *Mathematics of Operations Research* 17: 327–340.

Battigalli, P., and G. Bonanno. 1999. Recent results on belief, knowledge and the epistemic foundations of game theory. *Research in Economics* 53: 149–225.

Ben-Or, M. 1983. Another advantage of free choice: Completely asynchronous agreement protocols. In *Proceedings of the second ACM symposium on principles of distributed computing*, 27–30.

Ben-Porath, E. 2003. Cheap talk in games with incomplete information. *Journal of Economic Theory* 108: 45–71.

Blum, A., and Y. Mansour. 2007. Learning, regret minimization, and equilibria. In *Algorithmic game theory*, ed. N. Nisan, T. Roughgarden, É. Tardos, and V. Vazirani. Cambridge: Cambridge University Press.

Blum, B., C.R. Shelton, and D. Koller. 2003. A continuation method for Nash equilibria in structured games. In *Proceedings of the 18th international joint conference on artificial intelligence* (*IJCAI '03*), 757–764.

Brafman, R.I., and M. Tennenholtz. 2002. R-MAX: A general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research* 3: 213–231.

Chen, X., and X. Deng. 2006. Settling the complexity of 2-player Nash equilibrium. In *Proceedings of the 47th IEEE symposium on foundations of computer science*, 261–272.

Chor, B., and C. Dwork. 1989. Randomization in Byzantine agreement. In *Advances in computing research 5: Randomness and computation*, ed. S. Micali, 443–497. Greenwich: JAI Press.

Chu, F., and J.Y. Halpern. 2001. A decision-theoretic approach to reliable message delivery. *Distributed Computing* 14: 359–389.

Clarke, E.H. 1971. Multipart pricing of public goods. *Public Choice* 11: 17–33.

Conitzer, V., and T. Sandholm. 2003. Complexity results about Nash equilibria. In *Proceedings of the 18th international joint conference on artificial intelligence* (*IJCAI '03*), 765–771.

Conitzer, V., T. Sandholm, and J. Lang. 2007. When are elections with few candidates hard to manipulate? *Journal of the ACM* 54(3), Article 14.

Cramton, P., Y. Shoham, and R. Steinberg. 2006. *Combinatorial auctions*. Cambridge, MA: MIT Press.

Daskalis, C., P.W. Goldberg, and C.H. Papadimitriou. 2006. The complexity of computing a Nash equilibrium. In *Proceedings of the 38th ACM symposium on theory of computing*, 71–78.

Dekel, E., B. Lipman, and A. Rustichini. 1998. Standard state-space models preclude unawareness. *Econometrica* 66: 159–173.

Eliaz, K. 2002. Fault-tolerant implementation. *Review of Economic Studies* 69: 589–610.

Fabrikant, A., A. Luthra, E. Maneva, C.H. Papadimitriou, and S. Shenker. 2003. On a network creation game. In *Proceedings of the 22nd ACM symposium on principles of distributed computing*, 347–351.

Fagin, R., and J.Y. Halpern. 1988. Belief, awareness, and limited reasoning. *Artificial Intelligence* 34: 39–76.

Fagin, R., J.Y. Halpern, Y. Moses, and M.Y. Vardi. 1995. *Reasoning about knowledge*. Cambridge, MA: MIT Press A slightly revised paperback version was published in 2003.

Feigenbaum, J., C. Papadimitriou, and S. Shenker. 2000. Sharing the cost of muliticast transmissions (preliminary version). In *Proceedings of the 32nd ACM symposium on theory of computing*, 218–227.

Fischer, M.J. 1983. The consensus problem in unreliable distributed systems. In *Foundations of computation theory*, vol. 185, ed. M. Karpinski. Lecture Notes in Computer Science. Berlin/New York: Springer, pp. 127–140.

Fischer, M.J., N.A. Lynch, and M.S. Paterson. 1985. Impossibility of distributed consensus with one faulty processor. *Journal of the ACM* 32: 374–382.

Forges, F. 1990. Universal mechanisms. *Econometrica* 58: 1341–1364.

Fudenberg, D., and D. Levine. 1998. *The theory of learning in games*. Cambridge, MA: MIT Press.

Gerardi, D. 2004. Unmediated communication in games with complete and incomplete information. *Journal of Economic Theory* 114: 104–131.

Gibbard, A. 1973. Manipulation of voting schemes. *Econometrica* 41: 587–602.

Gilboa, I., and E. Zemel. 1989. Nash and correlated equilibrium: Some complexity considerations. *Games and Economic Behavior* 1: 80–93.

Goldreich, O., S. Micali, and A. Wigderson. 1987. How to play any mental game. In *Proceedings of the 19th ACM symposium on theory of computing*, 218–229.

Govindan, S., and R. Wilson. 2003. A global Newton method to compute Nash equilibria. *Journal of Economic Theory* 110: 65–86.

Groves, T. 1973. Incentives in teams. *Econometrica* 41: 617–631.

Halldórsson, M.M., J.Y. Halpern, L. Li, and V. Mirrokni. 2004. On spectrum sharing games. In *Proceedings of the 23rd ACM symposium on principles of distributed computing*, 107–114.

Halpern, J.Y. 2001. Alternative semantics for unawareness. *Games and Economic Behavior* 37: 321–339.

Halpern, J.Y. 2003. A computer scientist looks at game theory. *Games and Economic Behavior* 45: 114–132.

Halpern, J.Y., and L.C. Rêgo. 2007. Reasoning about knowledge of unawareness. *Games and Economic Behavior.* Also available at http://arxiv.org/abs/cs.LO/0603020. Accessed 24 June 2007.

Halpern, J.Y., and V. Teague. 2004. Rational secret sharing and multiparty computation: Extended abstract. In *Proceedings of the 36th ACM symposium on theory of computing*, 623–632.

Heifetz, A., M. Meier, and B. Schipper. 2006. Interactive unawareness. *Journal of Economic Theory* 130: 78–94.

Heller, Y. 2005. *A minority-proof cheap-talk protocol*. Unpublished manuscript.

Hopcroft, J.E., and J.D. Ullman. 1979. *Introduction to automata theory, languages and computation*. New York: Addison-Wesley.

Hurwicz, L. 1977. On the dimensional requirements of informationally decentralized Pareto satisfactory processes. In *Studies in Resource Allocation Processes*, ed. K.J. Arrow and L. Hurwicz, 413–424. New York: Cambridge University Press.

Izmalkov, S., S. Micali, and M. Lepinski. 2005. Rational secure computation and ideal mechanism design. In *Proceedings of the 46th IEEE symposium foundations of computer science*, 585–595.

Kearns, M. 2007. Graphical games. In *Algorithmic game theory*, ed. N. Nisan, T. Roughgarden, É. Tardos, and V. Vazirani. Cambridge: Cambridge University Press.

Kearns, M.J., and M.K. Reiter. 2005. *Proceedings of the sixth ACM conference on electronic commerce* (*EC '05*). New York: ACM. Table of contents available at http://www.informatik.uni-trier.de/~ley/db/conf/sigecom/sigecom2005.html. Accessed 19 June 2007.

Kearns, M., and S.P. Singh. 1998. Near-optimal reinforcement learning in polynomial time. In *Proceedings of the 15th international conference on machine learning*, 260–268.

Kearns, M., M.L. Littman, and S.P. Singh. 2001. Graphical models for game theory. In *Proceedings of the 17th conference on uncertainty in artificial intelligence* (*UAI 2001*), 253–260.

Kempe, D., J. Kleinberg, and É. Tardos. 2003. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on knowledge discovery and data mining*, 137–146.

Kfir-Dahav, N.E., D. Monderer, and M. Tennenholtz 2000. Mechanism design for resource-bounded agents. In *International conference on multiagent systems*, 309–316.

Koller, D., and B. Milch. 2001. Structured models for multiagent interactions. In *Theoretical aspects of rationality and knowledge: Proceedings of the eighth conference* (*TARK 2001*), 233–248.

Koutsoupias, E., and C.H. Papadimitriou. 1999. Worst-case equilibria. In *Proceedings of the 16th conference on theoretical aspects of computer science*, vol. 1563, Lecture Notes in Computer Science. Berlin: Springer, pp. 404–413.

Kushilevitz, E., and N. Nisan. 1997. *Communication complexity*. Cambridge: Cambridge University Press.

La Mura, P. 2000. Game networks. In *Proceedings of the 16th conference on uncertainty in artificial intelligence* (*UAI2000*), 335–342.

Lehmann, D., R. Müller, and T. Sandholm. 2006. The winner determination problem. In *Combinatorial auctions*, ed. P. Cramton, Y. Shoham, and R. Steinberg. Cambridge, MA: MIT Press.

Lemke, C.E., and J.J.T. Howson. 1964. Equilibrium points of bimatrix games. *Journal of the Society for Industrial and Applied Mathematics* 12: 413–423.

Lepinski, M., S. Micali, C. Peikert, and A. Shelat. 2004. Completely fair SFE and coalition-safe cheap talk. In *Proceedings of the 23rd ACM symposium principles of distributed computing*, 1–10.

Linial, N. 1994. Games computers play: Game-theoretic aspects of computing. In *Handbook of game theory*, ed. R.J. Aumann and S. Hart, vol. 2. Amsterdam: North-Holland.

Modica, S., and A. Rustichini. 1994. Awareness and partitional information structures. *Theory and Decision* 37: 107–124.

Modica, S., and A. Rustichini. 1999. Unawareness and partitional information structures. *Games and Economic Behavior* 27: 265–298.

Monderer, D., and M. Tennenholtz. 1999a. Distributed games. *Games and Economic Behavior* 28: 55–72.

Monderer, D., and M. Tennenholtz. 1999b. Distributed games: From mechanisms to protocols. In *Proceedings of the 16th national conference on artificial intelligence* (*AAAI '99*), 32–37.

Moreno, D., and J. Wooders. 1996. Coalition-proof equilibrium. *Games and Economic Behavior* 17: 80–112.

Mount, K., and S. Reiter. 1974. The informational size of message spaces. *Journal of Economic Theory* 8: 161–192.

Nash, J. 1950. Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences* 36: 48–49.

Neyman, A. 1985. Bounded complexity justifies cooperation in finitely repeated Prisoner's Dilemma. *Economic Letters* 19: 227–229.

Nisan, N. 2006. Bidding languages for combinatorial auctions. In *Combinatorial auctions*. Cambridge, MA: MIT Press.

Nisan, N., and A. Ronen. 2000. Computationally feasible VCG mechanisms. In *Proceedings of the second ACM conference on electronic commerce* (*EC '00*), 242–252.

Nisan, N., and A. Ronen. 2001. Algorithmic mechanism design. *Games and Economic Behavior* 35: 166–196.

Nisan, N., T. Roughgarden, É. Tardos, and V. Vazirani. 2007. *Algorithmic game theory*. Cambridge: Cambridge University Press.

Papadimitriou, C.H. 1994a. *Computational complexity*. Reading: Adison-Wesley.

Papadimitriou, C.H. 1994b. On the complexity of the parity argument and other inefficient proofs of existence. *Journal of Computer and System Sciences* 48: 498–532.

Papadimitriou, C.H. 2001. Algorithms, games, and the internet. In *Proceedings of the 33rd ACM symposium on theory of computing*, 749–753.

Papadimitriou, C.H. 2005. Computing correlated equilibria in multiplayer games. In *Proceedings of the 37th ACM symposium on theory of computing*, 49–56.

Papadimitriou, C.H. 2007. The complexity of finding Nash equilibria. In *Algorithmic game theory*, ed. N. Nisan, T. Roughgarden, É. Tardos, and V. Vazirani. Cambridge: Cambridge University Press.

Papadimitriou, C.H., and T. Roughgarden. 2005. Computing equilibria in multi-player games. In *Proceedings of the 16th ACM–SIAM symposium on discrete algorithms*, 82–91.

Papadimitriou, C.H., and M. Yannakakis. 1994. On complexity as bounded rationality. In *Proceedings of the 26th ACM symposium on theory of computing*, 726–733.

Pearl, J. 1988. *Probabilistic reasoning in intelligent systems*. San Francisco: Morgan Kaufmann.

Pease, M., R. Shostak, and L. Lamport. 1980. Reaching agreement in the presence of faults. *Journal of the ACM* 27: 228–234.

Porter, R., E. Nudelman, and Y. Shoham. 2004. Simple search methods for finding a Nash equilibrium. In *Proceedings of the 21st national conference on artificial intelligence* (*AAAI '04*), 664–669.

Rabin, M.O. 1983. Randomized Byzantine generals. In *Proceedings of the 24th IEEE symposium on foundations of computer science*, 403–409.

Roughgarden, T., and É. Tardos. 2002. How bad is selfish routing? *Journal of the ACM* 49: 236–259.

Rubinstein, A. 1986. Finite automata play the repeated prisoner's dilemma. *Journal of Economic Theory* 39: 83–96.

Rubinstein, A. 1998. *Modeling bounded rationality*. Cambridge, MA: MIT Press.

Sandholm, T., and Yokoo, M. 2003. *The second international joint conference on Autonomous Agents and Multiagent Systems* (*AAMAS 2003*). Table of contents available at http://www.informatik.uni-trier.de/~ley/db/conf/atal/aamas2003.html. Accessed 25 June 2007.

Satterthwaite, M. 1975. Strategy-proofness and Arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of Economic Theory* 10: 187–217.

Segal, I. 2006. Communication in economic mechanisms. In *Advances in economics and econometrics: Theory and application, ninth world congress (Econometric society monographs)*, ed. R. Blundell, W.K. Newey, and T. Persson, 222–268. Cambridge: Cambridge University Press.

Shamir, A., R.L. Rivest, and L. Adelman. 1981. Mental poker. In *The mathematical gardner*, ed. D.A. Klarner, 37–43. Boston: Prindle, Weber, and Schmidt.

Tennenholtz, M. 2002. Competitive safety analysis: Robust decision-making in multi-agent systems. *Journal of Artificial Intelligence Research* 17: 363–378.

Urbano, A., and J.E. Vila. 2002. Computational complexity and communication: Coordination in two-player games. *Econometrica* 70: 1893–1927.

Urbano, A., and J.E. Vila. 2004. Computationally restricted unmediated talk under incomplete information. *Economic Theory* 23: 283–320.

Vetta, A. 2002. Nash equilibria in competitive societies, with applications to facility location, traffic routing and auctions. In *Proceedings of the 43rd IEEE symposium on foundations of computer science*, 416–425.

Vickrey, W. 1961. Counterspeculation, auctions and competitive sealed tenders. *Journal of Finance* 16: 8–37.

Yao, A. 1982. Protocols for secure computation (extended abstract). In *Proceedings of the 23rd IEEE symposium on foundations of computer science*, 160–164.

Computing in Mechanism Design

Tuomas Sandholm

Abstract

Computational issues are important in mechanism design, but have received insufficient research interest. This article briefly reviews some of the key ideas. I discuss computing by the *centre*, such as an auction server or vote aggregator, and computing by the *agents*, be they human or software. Limited computing hinders mechanism design in several ways, and presents deep strategic interactions between computing and incentives. On the bright side, novel algorithms and increasing computing power have enabled better mechanisms. Perhaps most interestingly, with computationally limited agents, one can implement mechanisms that would not be implementable among computationally unlimited agents.

Keywords

Algorithmic mechanism design; Automated mechanism design; Borda voting rule; Bounded rationality; Combinatorial auctions; Complexity theory; Computing by the centre; Computing in mechanism design; Deliberation equilibrium; Elicitor; *ex post* equilibrium; Expressive commerce; Fielded combinatorial auctions; Fielded expressive auctions; Gibbard–Satterthwaite th; Incentive compatibility; Maximin voting rule; Non-truth-promoting mechanism; Mechanism design; Performance profile tree; Plurality voting rule; Preference determination; Preference elicitation; Pull–pushmechanism; Revelation principle; Strategic computing; Tree search; Vickrey auction; Vickrey–Clarke–Groves (VCG) mechanism

JEL Classifications

C7

Introduction

Computational issues in mechanism design are important, but have received insufficient research interest until recently. Limited computing hinders mechanism design in several ways, and presents deep strategic interactions between computing and incentives. On the bright side, novel algorithms and increasing computing power have enabled better mechanisms. Perhaps most interestingly, limited computing of the agents can be used as a tool to implement mechanisms that would not be implementable among computationally unlimited agents. This article briefly reviews some of the key ideas, with the goal of alerting the reader to the importance of these issues and hopefully spurring future research.

I will discuss computing by the *centre*, such as an auction server or vote aggregator, in Section "Computing by the Centre". Then, in Section "Computing by the Agents", I will address the *agents*' computing, be they human or software.

Computing by the Centre

Computing by the centre plays significant roles in mechanism design. In the following three subsections I will review three prominent directions.

Executing Expressive Mechanisms

As algorithms have advanced drastically and computing power has increased, it has become feasible to field mechanisms that were previously impractical. The most famous example is a *combinatorial auction* (*CA*). In a CA, there are multiple distinguishable items for sale, and the bidders can submit bids on self-selected packages of the items. (Sometimes each bidder is also allowed to submit exclusivity constraints of different forms among his bids.) This increase in the expressiveness of the bids drastically reduces the strategic complexity that bidders face. For one, it removes the exposure problems that bidders face when they have preferences over packages but in traditional auctions are allowed to submit bids on individual items only.

CAs shift the computational burden from the bidders to the centre. There is an associated gain because the centre has all the information in hand to optimize while in traditional auctions the bidders only have estimated projected (probabilistic) information about how others will bid. Thus CAs yield more efficient allocations.

On the downside, the centre's task of determining the winners in a CA (deciding which bids to accept so as to maximize the sum of the accepted bids'prices subject to not selling any item to more than one bid) is a complex combinatorial optimization problem, even without exclusivity constraints among bids. Three main approaches have been studied for solving it.

1. *Optimal winner determination using some form of tree search*. For a review, see Sandholm (2006). The advantage is that the bidding language is not restricted and the optimal solution is found. The downside is that no optimal winner determination algorithm can run in polynomial time in the size of the problem instance in the worst case, because the problem is $\mathcal{NP}$-complete (Rothkopf et al. 1998). ($\mathcal{NP}$-complete problems are problems for which the fastest known algorithms take exponential time in the size of the problem instance in the worst case. $\mathcal{P}$ is the class of easy problems solvable in polynomial time. The statement of winner determination not being solvable in polynomial time in the worst case relies on the usual assumption $\mathcal{P} \neq \mathcal{NP}$. This is an open question in complexity theory, but is widely believed to be true. If false, that would have sweeping implications throughout computer science.)
2. *Approximate winner determination*. The advantage is that many approximation algorithms run in polynomial time in the size of the instance even in the worst case. For reviews of such algorithms, see Sandholm (2002a) and Lehmann et al. (2006). (Other suboptimal algorithms do not have such time guarantees, such as local search, stochastic local search, simulated annealing, genetic algorithms and tabu search.) The downside is that the solution is sometimes far from optimal: no such algorithm can always find a solution that is within a factor

$$\min\left\{\#bids^{1-\varepsilon},\ items^{\frac{1}{2}-\varepsilon}\right\} \tag{1}$$

of optimal (Sandholm 2002a). (This assumes $\mathcal{LPP} \neq \mathcal{NP}$. It is widely believed that these two complexity classes are indeed unequal.) For example, with just nine items for sale, no such algorithm can extract even 33 per cent of the available revenue from the bids in the worst case. With 81 items, that drops to 11 per cent.
3. *Restricting the bidding language* so much that optimal (within the restricted language) winner determination can be conducted in worst-case polynomial time. For a review, see Müller (2006). For example, if each package bid is only allowed to include at most two items, then winners can be determined in worst-case polynomial time (Rothkopf et al. 1998). The downside is that bidders have to shoehorn their preferences into a restricted bidding language; this gives rise to similar problems as in non-combinatorial mechanisms for multi-item auctions: exposure problems, need to speculate how others will bid, inefficient allocation, and so on.

Truthful bidding can be made a dominant strategy by applying the *Vickrey–Clarke–Groves* (*VCG*) *mechanism* to a CA. Such incentive compatibility removes strategic complexity of the bidders. The mechanism works as follows. The optimal allocation is used, but the bidders do not pay their winning bids. Instead each bidder pays the amount of value he takes away from the others by taking some of the items. This value is measured as the difference between the others' winning bids' prices and what the others' winning bids' prices would have been had the agent not submitted any bids. This mechanism can be executed by determining the winners once overall, and once for each agent removed in turn. (This may be accomplishable with less computing. For example, in certain network auctions it can be done in the same asymptotic complexity as one winner determination – Hershberger and Suri 2001.)

Very few canonical CAs have found their way to practice. However, auctions with richer bid expressiveness forms (that are more natural in the given application and more concise) and that support expressiveness also by the bid taker have made a major breakthrough into practice (Sandholm 2007; Bichler et al. 2006). This is sometimes called *expressive commerce* to distinguish it from vanilla CAs. The widest area of application is currently industrial sourcing. Tens of billions of dollars worth of materials, transportation and services are being sourced annually using such mechanisms, yielding billions of dollars in efficiency improvements. The bidders' expressiveness forms include different forms of flexible package bids, conditional discounts, discount schedules, side constraints (such as capacity constraints), and often hundreds of cost drivers (for example, fixed costs, variable costs, transshipment costs and costs associated with changes). The item specifications can also be left partially open, and the bidders can specify some of the item attributes (delivery date, insurance terms, and so on). in alternate ways. The bid taker also specifies preferences and constraints. Winner determination then not only decides who wins what, but also automatically configures the items. In some of these events it also configures the supply chain several levels deep as a side effect. On the high end, such an auction can have tens of thousands of items (multiple units of each), millions of bids, and hundreds of thousands of side constraints. Expressive mechanisms have also been designed for settings beyond auctions, such as combinatorial exchanges, charity donations and settings with externalities.

Basically all of the fielded expressive auctions use the simple pay-your-winning-bids pricing rule. There are numerous important reasons why few, if any, use the VCG mechanism. It can lead to low revenue. It is vulnerable to collusion. Bidders would not tell the truth because they do not want to reveal their cost structures, which the auctioneer could exploit the next time the auction is conducted, and so on (Sandholm 2000; Rothkopf 2007).

Basically all of the fielded expressive auctions use tree search for winner determination. In practice, modern tree search algorithms for the problem scale to the large and winners can be determined optimally. If winner determination were not done optimally in a CA, the VCG mechanism can lose its truth-dominance property (Sandholm 2002b). In fact, any truthful suboptimal VCG-based mechanism for CAs is unreasonable in the sense that it sometimes does not allocate an item to a bidder even if he is the only bidder whose bids assign non-zero value to that item (Nisan and Ronen 2000).

Algorithmic Mechanism Design

Motivated by the worry that some instances of *NP*-hard problems may not be solvable within reasonable time, a common research direction in theory of computing is approximation algorithms. They trade off solution quality for a guarantee that even in the worst case, the algorithm runs in polynomial time in the size of the input.

Analogously, Nisan and Ronen (2001) proposed *algorithmic mechanism design*: designing approximately optimal mechanisms that take the centre a polynomial number of computing steps even in the worst case. However, this is more difficult than designing approximately optimal algorithms because the mechanism has to motivate the agents to tell the truth.

Lehmann et al. (2002) studied this for CAs with single-minded bidders (each bidder being only interested in one specific package of items). They present a fast greedy algorithm that guarantees a solution within a factor $\sqrt{\#items}$ of optimal. They show that the algorithm is not incentive compatible with VCG pricing, but is with their custom pricing scheme. They also identify sufficient conditions for any (approximate) mechanism to be incentive compatible (see also Kfir-Dahav et al. 2000). There has been substantial follow-on work on subclasses of single-minded CAs.

Lavi and Swamy (2005) developed a technique for a range of packing problems with which any k-approximation algorithm (that is, algorithm that guarantees that the solution is within a factor k of optimal) that also bounds the integrality gap of the linear programming (LP) relaxation of the problem by k can be used to construct a $k$approximation mechanism. The LP solution, scaled down by k, can be represented as a convex combination of integer solutions, and viewing this convex combination as specifying a probability distribution over integer solutions begets a VCG-based randomized mechanism that is truthful in expectation. For CAs with general valuations, this yields an $O\left(\sqrt{\#items}\right)$ -approximate mechanism.

In a different direction, several mechanisms have been proposed where the agents can help the centre find better outcomes. This is done either by giving the agents the information to do the centre's computing (Banks et al. 1989; Land et al. 2006; Parkes and Shneidman 2004), or by allowing the agents to change what they told the mechanism based on the mechanism's output and potentially also based on what other agents told the mechanism (Nisan and Ronen 2000). In VCG-based mechanisms, an agent benefits from lying only if the lie causes the mechanism to find an outcome that is better overall.

Automated Mechanism Design

Conitzer and Sandholm (2002) proposed the idea of *automated mechanism design*: having a computer, rather than a human, design the mechanism. Because human effort is eliminated, this enables custom design of mechanisms for every setting. The setting can be described by the agents' (discretized) type spaces, the designer's prior over types, the desired notion of incentive compatibility (for example, dominant strategies vs. Bayes–Nash implementation), the desired notion of participation constraints (for example, *ex interim*, *ex post* or none), whether payments are allowed, and whether the mechanism is allowed to use randomization.) This can yield better mechanisms for previously studied settings because the mechanism is designed for the specific setting rather than a class of settings. It can also be used for settings not previously studied in mechanism design.

For almost all natural (linear) objectives, all variants of the design problem are $\mathcal{NP}$-complete if the mechanism is not allowed to use randomization, but randomized mechanisms can be constructed for all these settings in polynomial time using linear programming. Custom algorithms have been developed for some problems in each of these two categories. (Even the latter category warrants research. While the linear programme is polynomial in the size of the input, the input itself can be exponential in the number of agents.) Structured representations of the problem can also make the design process drastically faster.

Beyond the general setting, automated mechanism design has been applied to specific settings, such as creating revenue-maximizing CAs (without the need to discretize types)–Likhodedov and Sandholm 2005 (a recognized problem that eludes analytical characterization; even the two-item case is open), reputation systems (Jurca and Faltings 2006), safe exchange mechanisms (Sandholm and Ferrandon 2000), and supply chain settings (Vorobeychik et al. 2006). Automated mechanism design software has recently also been adopted by several mechanism design theoreticians to speed up their research.

It turns out that even *multistage mechanisms* can be designed automatically (Sandholm et al. 2007). Furthermore, automated mechanism design has been applied to the design of *online mechanisms* (Hajiaghayi et al. 2007), that is,

mechanisms that execute while the world changes –for example, agents enter and exit the system.

Computing by the Agents

I will now move to discussing computing by the agents.

Mechanisms That Are Hard to Manipulate

This section demonstrates that one can use the fact that agents are computationally limited to achieve things that are not achievable via any mechanism among perfectly rational agents.

A seminal negative result, the *Gibbard–Satterthwaite th*, states that if there are three or more candidates, then in any non-dictatorial voting scheme there are candidate rankings of the other voters, and preferences of the agent, under which the agent is better off voting manipulatively than truthfully. One avenue around this impossibility is to construct desirable general non-dictatorial voting protocols under which *finding* a beneficial manipulation is prohibitively hard computationally.

There are two natural alternative goals of manipulation. In *constructive manipulation*, the manipulator tries to find an order of candidates that he can reveal so that his favourite candidate wins. In *destructive manipulation*, the manipulator tries to find an order of candidates that he can reveal so that his hated candidate does not win. These are special cases of the utility-theoretic notion of improving one's utility, so the hardness results, discussed below, carry over to the usual utility-theoretic setting.

Unfortunately, finding a constructive manipulation is easy (in $\mathcal{P}$) for the *plurality*, *Borda* and *maximin* voting rules (Bartholdi et al. 1989), which are commonly used. On the bright side, constructive manipulation of the *single transferable vote* (*STV*) protocol is $\mathcal{NP}$-hard (Bartholdi and Orlin 1991) (as is manipulation of the *second order Copeland* protocol (Bartholdi et al. 1989), but that hardness is driven solely by the tie-breaking rule). Even better, there is a systematic methodology for slightly tweaking voting protocols that are easy to manipulate, so that they become hard to manipulate (Conitzer and Sandholm 2003). Specifically, before the original protocol is executed, one pairwise elimination round is executed among the candidates, and only the winning candidates survive to the original protocol. This makes the protocols $\mathcal{NP}$-hard, # $\mathcal{P}$hard (#$\mathcal{P}$-hard problems are at least as hard as counting the number of solutions to a problem in $\mathcal{P}$), or even *PSPACE*-hard (*PSPACE*-hard problems are at least as hard as any problem that can be solved using a polynomial amount of memory) to manipulate constructively, depending on whether the schedule of the pre-round is determined before the votes are collected, randomly after the votes are collected, or the scheduling and the vote collecting are carefully interleaved, respectively.

All of the hardness results of the previous paragraph rely on both the number of voters and the number of candidates growing. The number of candidates can be large in some domains, for example when voting over task or resource allocations.

However, in other elections – such as presidential elections – the number of candidates is small. If the number of candidates is a constant, both constructive and destructive manipulation are easy (in $\mathcal{P}$), regardless of the number of voters (Conitzer et al. 2007). This holds even if the voters are weighted, or if a coalition of voters tries to manipulate. On the bright side, when a coalition of weighted voters tries to manipulate, complexity can arise even for a constant number of candidates: see Tables 1 and 2. Another lesson from that table is that randomizing over instantiations of the mechanism (such as schedules of a *cup*) can be used to make manipulation hard.

As usual in computer science, all the results mentioned above are worst-case hardness. Unfortunately, under weak assumptions on the preference distribution and voting rule, most instances of any voting rule are easy to manipulate (Conitzer and Sandholm 2006).

All of the hardness results discussed above hold even if the manipulators know the non-manipulators' votes exactly. Under weak

Computing in Mechanism Design, Table 1 Complexity of constructive weighted coalitional manipulation

Number of candidates:	2	3	4, 5, 6	≥7
Borda	$\mathcal{P}$	$\mathcal{NP}$ -complete	$\mathcal{NP}$-complete	$\mathcal{NP}$-complete
Veto	$\mathcal{P}$	$\mathcal{NP}$-complete	$\mathcal{NP}$-complete	$\mathcal{NP}$-complete
STV	$\mathcal{P}$	$\mathcal{NP}$-complete	$\mathcal{NP}$-complete	$\mathcal{NP}$-complete
Plurality with runoff	$\mathcal{P}$	$\mathcal{NP}$-complete	$\mathcal{NP}$-complete	$\mathcal{NP}$-complete
Copeland	$\mathcal{P}$	$\mathcal{P}$	$\mathcal{NP}$-complete	$\mathcal{NP}$-complete
Maximin	$\mathcal{P}$	$\mathcal{P}$	$\mathcal{NP}$-complete	$\mathcal{NP}$-complete
Randomized cup	$\mathcal{P}$	$\mathcal{P}$	$\mathcal{P}$	$\mathcal{NP}$-complete
Cup	$\mathcal{P}$	$\mathcal{P}$	$\mathcal{P}$	$\mathcal{P}$
Plurality	$\mathcal{P}$	$\mathcal{P}$	$\mathcal{P}$	$\mathcal{P}$

Computing in Mechanism Design, Table 2 Complexity of destructive weighted coalitional manipulation

Number of candidates:	2	≥3
STV	$\mathcal{P}$	$\mathcal{NP}$-complete
Plurality with runoff	$\mathcal{P}$	$\mathcal{NP}$-complete
Randomized cup	$\mathcal{P}$	?
Borda	$\mathcal{P}$	$\mathcal{P}$
Veto	$\mathcal{P}$	$\mathcal{P}$
Copeland	$\mathcal{P}$	$\mathcal{P}$
Maximin	$\mathcal{P}$	$\mathcal{P}$
Cup	$\mathcal{P}$	$\mathcal{P}$
Plurality	$\mathcal{P}$	$\mathcal{P}$

Source: Conitzer et al. (2007)

assumptions, if weighted coalitional manipulation with complete information about the others' votes is hard in some voting protocol, then individual unweighted manipulation is hard when there is uncertainty about the others' votes (Conitzer et al. 2007).

Non-Truth-Promoting Mechanisms

A challenging issue is that even if it is prohibitively hard to find a beneficial manipulation, the agents might not tell the truth. For example, an agent might take a chance that he will do better with a lie. The following result shows that, nevertheless, mechanism design can be improved by making the agents face complexity. (This is one reason why computational issues can render the *revelation principle* inapplicable. One of the things the principle says is that for any non-truth-promoting mechanism it is possible to construct an incentive-compatible mechanism that is at least as good. The theorem below challenges this.)

Theorem 1 (Conitzer and Sandholm 2004) *Suppose the centre is trying to maximize social welfare, and neither payments nor randomization is allowed. Then, even with just two agents (one of whom does not even report a type, so dominant strategy implementation and Bayes–Nash implementation coincide), there exists a family of preference aggregation settings such that:*

- *the execution of any optimal incentive-compatible mechanism is $\mathcal{NP}$ -complete for the center, and.*
- *there exists a non-incentive-compatible mechanism which (1) requires the centre to carry out only polynomial computation, and (2) makes finding any beneficial insincere revelation NP-complete for the type-reporting agent. Additionally, if the type-reporting agent manages to find a beneficial insincere revelation, or no beneficial insincere revelation exists, the social welfare of the outcome is identical to the social welfare that would be produced by*

any optimal incentive-compatible mechanism. Finally, if the type-reporting agent does not manage to find a beneficial insincere revelation where one exists, the ***social welfare of the outcome is strictly greater than the social welfare that would be produced by any optimal incentive-compatible mechanism***.

An analogous theorem holds if, instead of counting computational steps, we count calls to a commonly accessible oracle which, when supplied with an agent, that agent's type, and an outcome, returns a utility value for that agent.

Preference (Valuation) Determination Via Computing or Information Acquisition

In many (auction) settings, even determining one's valuation for an item (or a bundle of items) is complex. For example, when bidding for trucking lanes (tasks), this involves solving two $\mathcal{NP}$-complete local planning problems: the vehicle routing problem with the new lanes of the bundle and the problem without them (Sandholm 1993). The difference in the costs of those two local plans is the cost (valuation) of taking on the new lanes.

In these types of settings, the *revelation principle* applies only in a trivial way: the agents report their data and optimization models to the centre, and the centre does the computation for them. It stands to reason that in many applications the centre would not want to take on that burden, in which case such extreme direct mechanisms are not an option. Therefore, I will now focus on mechanisms where the agents report valuations to the centre, as in traditional auctions.

Bidders usually have limited computing and time, so they cannot exactly evaluate all (or even any) bundles – at least not without cost. This leads to a host of interesting issues where computing and incentives are intimately intertwined.

For example, in a one-object auction, should a bidder evaluate the object if there is a cost to doing so? It turns out that the Vickrey auction loses its dominantstrategy property: whether or not the bidder should pay the evaluation cost depends on the other bidders' valuations (Sandholm 2000).

If a bidder has the opportunity to *approximate his valuation to different degrees*, how much computing time should the bidder spend on refining its valuation? If there are multiple items for sale, how much computing time should the bidder allocate on different bundles? A bidder may even allocate some computing time to evaluate other bidders' valuations so as to be able to bid more strategically; this is called *strategic computing*.

To answer these qsts, Larson and Sandholm (2001) developed a deliberation control method called a *performance profile tree* for projecting how an anytime algorithm (that is, an algorithm that has an answer available at any time, but where the quality of the answer improves the more computing time is allocated to the algorithm) will change the valuation if additional computing is allocated toward refining (or improving) it. This deliberation control method applies to any anytime algorithm. Unlike earlier deliberation control methods for anytime algorithms, the performance profile tree is a *fully normative model of bounded rationality*: it takes into account all the information that an agent can use to make its deliberation control decisions. This is necessary in the game-theoretic context; otherwise a strategic agent could take into account some information that the model does not.

Using this deliberation control method, the auction can be modelled as a game where the agents' strategy spaces include computing actions. At every point, each agent can decide on which bundle to allocate its next step of computing as a function of the agent's computing results so far (and in open-cry auction format also the others' bids observed so far). At every point, the agent can also decide to submit bids. One can then solve this for equilibrium: each agent's (deliberation and bidding) strategy is a best-response to the others' strategies. This is called *deliberation equilibrium*.

This notion, and the performance profile tree, apply not only to computational actions but also to information gathering actions for determining valuations. (In contrast, most of the literature on information acquisition in auctions does not take

Computing in Mechanism Design, Table 3 Can strategic computing occur in deliberation equilibrium? The most interesting results are in bold. As a benchmark from classical auction theory, the table also shows whether or not perfectly rational agents, that can determine their valuations instantly without cost, would benefit from considering each others' valuations when deciding how to bid

	Auctionmechanism	Speculation by perfectly rational agents?	Strategic computing	
			Limited computing	**Costly computing**
Single item	First price	Yes	Yes	Yes
	Dutch	Yes	Yes	Yes
	English	No	**No**	**Yes**
	Vickrey	No	**No**	**Yes**
Multiple items	First price	Yes	Yes	Yes
	VCG	No	**Yes**	**Yes**

into account that valuations can be determined to different degrees and that an agent may want to invest effort to determine others' valuations as well – even in privatevalue settings.)

Table 3 shows in which settings strategic computing can and cannot occur in deliberation equilibrium. This depends on the auction mechanism. Interestingly, it also depends on whether the agent has limited computing (for example, owning a desktop computer that the agent can use until the auction's deadline) or costly computing (for example, being able to buy any amount of supercomputer time where each cycle comes at a cost).

The notion of deliberation equilibrium can also be used as the basis for designing new mechanisms, which hopefully work well among agents whose computing is costly or limited. Unfortunately, there is an impossibility (Larson and Sandholm 2005): there exists no mechanism that is *sensitive* (the outcome is affected by each agent's strategy), *preference formation independent* (does not do the computations for the agents; the agents report valuations), *non-misleading* (no agent acts in a way that causes others to believe his true type has zero probability), and *deliberation-proof* (no strategic computing occurs in equilibrium, that is, agents compute only on their own problems). Current work involves designing mechanisms that take part in preference formation in limited ways: for example, agents report their performance profile trees to the centre, which then coordinates the deliberations incrementally as agents report deliberation results. Current research also includes designing mechanisms where strategic computing occurs but its wastefulness is limited.

Preference Elicitation by the Centre

To reduce the agents' preference determination effort, Conen and Sandholm (2001) proposed a framework where the centre (also known as *elicitor*) explicitly elicits preference information from the agents incrementally on an as-needed basis by posing queries to the agents. The centre thereby builds a model of the agents'preferences, and decides what to ask, and from which agent, based on this model. Usually the process can be terminated with the provably correct outcome while requiring only a small portion of the agents' preferences to be determined. Multistage mechanisms can yield up to exponential savings in preference determination and communication effort the agents need to go through compared to single-stage mechanisms (Conitzer and Sandholm 2004).

The explicit preference elicitation framework was originally proposed for CAs (but the approach has since been used for other settings as well, such as voting). For general valuations, an exponential number of bits in the number of items for sale has to be communicated in the worst case

no matter what queries are used (Nisan and Segal 2006). However, experimentally only a small fraction of the preference information needs to be elicited before the provably optimal solution is found. Furthermore, for valuations that have certain types of structure, even the worst-case number of queries needed is small. Research has also been done on the relative power of different query types.

If enough information is elicited to also determine the VCG payments, and these are the payments charged to the bidders, answering the elicitor's queries truthfully is an *ex post* equilibrium (a strengthening of Nash equilibrium that does not rely on priors). (This assumes there is no explicit cost or limit to valuation determination; mechanisms have also been designed for settings where there is an explicit cost (Larson 2006).) This holds even if the agents are allowed to answer queries that the elicitor did not ask (for example, queries that are easy for the agent to answer and which the agent thinks will significantly advance the elicitation process). We thus have a *pull–push mechanism* where both the centre and the agents guide the preference revelation (and thus also the preference determination/ refinement by the agents). For a review, see Sandholm and Boutilier (2006). Ascending (combinatorial) auctions are an earlier special case, and have limited power compared to the general framework (Blumrosen and Nisan 2005).

Preference elicitation can sometimes be computationally complex for the centre. It can be complex to intelligently decide what to ask next, and from whom. It can also be complex to determine whether enough information has been elicited to determine the optimal outcome. Even if the elicitor knows that enough has been elicited, it can be complex to determine the outcome – for example, allocation of items to bidders in some CAs.

Distributed (Centre-Free) Mechanisms

Computer scientists often have a preference for distributed applications that do not have any centralized coordination point (centre). Depending on the application, the reasons for this preference may include avoiding a single vulnerable point of failure, distributing the computing effort (for computational efficiency or because the data is inherently distributed), and enhancing privacy. The preference carries over from traditional computer science applications to different forms of negotiation systems –for example, see Sandholm (1993) for an early distributed automated negotiation system for software agents.

Feigenbaum et al. (2005) have studied lowest-cost inter-domain routing on the Internet, modifying a distributed protocol so that the agents (routing domains) are motivated to report their true costs and the solution is found with minimal message passing. For a review of some other research topics in this space, see Feigenbaum and Shenker (2002).

One can go further by taking into account the fact that agents might not choose to follow the prescribed protocol. They may cheat not only on information-revelation actions, but also on message-passing and computational actions. Despite computation actions not being observable by others, an agent can be motivated to compute as prescribed by tasking at least one other agent with the same computation, and comparing the results (Sandholm et al. 1999). Careful problem partitioning can also be used to achieve the same outcome without redundancy by only requiring agents to perform computing and message passing tasks that are in their own interest (Parkes and Shneidman 2004). Shneidman and Parkes (2004) propose a general proof technique and instantiate it to provide a non-manipulable protocol for inter-domain routing. Monderer and Tennenholtz (1999) develop protocols for one-item auctions executed among agents on a communication network. The protocols motivate the agents to correctly reveal preferences and communicate. For the setting where agents with private utility functions have to agree on variable assignments subject to side constraints (for example, meeting scheduling), Petcu et al. (2006) developed a VCG-based distributed optimization protocol that finds the social welfare maximizing allocation

and each agent is motivated to follow the protocol in terms of all three types of action. The only centralized party needed is a bank that can extract payments from the agents.

Cryptography is a powerful tool for achieving privacy when trying to execute a mechanism in a distributed way without a centre, using private communication channels among the agents. Consider first the setting with passive adversaries, that is, agents that faithfully execute the specified distributed communication protocol, but who try to infer (at least something about) some agents' private information.

- If agents are computationally limited – for example, they are assumed to be unable to factor large numbers – then arbitrary functions can be computed while guaranteeing that each agent maintains his privacy (except, of course, to the extent that the answer of the computation says something about the inputs) (Goldreich et al. 1987). Thus the desire for privacy does not constrain what social choice functions can be implemented.
- In contrast, only very limited social choice functions can be computed privately among computationally unlimited agents. For example, when there are just two alternatives, every monotonic, non-dictatorial social choice function that can be privately computed is constant (Brandt and Sandholm 2005). With special structure in the preferences, this impossibility can sometimes be avoided. For example, with the standard model of quasi-linear utility, first-price auctions can be implemented privately; second-price (Vickrey) auctions with more than two bidders cannot (Brandt and Sandholm 2004).

A more general model is that of active adversaries who can execute the distributed communication protocol unfaithfully in a coordinated way. A more game-theoretic model is that of rational adversaries that are not passive, but not malicious either. For a brief overview of such work, see computer science and game theory.

Acknowledgment This work was funded by the National Science Foundation under ITR grant IIS0427858, and a Sloan Foundation Fellowship. I thank Felix Brandt, Christina Fong, Joe Halpern, and David Parkes for helpful comments.

C

Bibliography

Banks, J.S., J. Ledyard, and D. Porter. 1989. Allocating uncertain and unresponsive resources: An experimental approach. *RAND Journal of Economics* 20: 1–25.

Bartholdi, J. III, and J. Orlin. 1991. Single transferable vote resists strategic voting. *Social Choice and Welfare* 8: 341–354.

Bartholdi, J. III, C. Tovey, and M. Trick. 1989. The computational difficulty of manipulating an election. *Social Choice and Welfare* 6: 227–241.

Bichler, M., A. Davenport, G. Hohner, and J. Kalagnanam. 2006. Industrial procurement auctions. In *Combinatorial auctions*, ed. Cramton, Shoham, and Steinberg. Cambridge, MA: MIT Press.

Blumrosen, L., and N. Nisan. 2005. On the computational power of iterative auctions. In *Proceedings of the ACM conference on electronic commerce*, 29–43. Vancouver: ACM Press.

Brandt F, andSandholm T. 2004. (Im)possibility of unconditionally privacypreserving auctions. In *Proceedings of the international conference on autonomous agents and multi-agent systems*.810–17.

Brandt F., and T. Sandholm. 2005. Unconditional privacy in social choice. In *Proceedings of the conference on theoretical aspects of rationality and knowledge*. 207–18.

Conen W., and T. Sandholm. 2001. Preference elicitation in combinatorial auctions: Extended abstract. In *Proceedings of the ACM conference on electronic commerce*, 256–9. More detailed description of algorithmic aspects in *Proceedings of the IJCAI01 workshop on economic agents, models, and mechanisms*. 71–80.

Conitzer V., and T. Sandholm. 2002. Complexity of mechanism design. In *Proceedings of the conference on uncertainty in artificial intelligence*. 103–10.

Conitzer V., and T. Sandholm. 2003. Universal voting protocol tweaks to make manipulation hard. In *Proceedings of the international joint conference on artificial intelligence*. 781–8.

Conitzer V., and T. Sandholm. 2004. Computational criticisms of the revelation principle. In *Conference on logic and the foundations of game and decision theory*. Earlier versions: AMEC-03, EC-04.

Conitzer V., and T. Sandholm. 2006. Nonexistence of voting rules that are usually hard to manipulate. In *Proceedings of the national conference on artificial intelligence*.

Conitzer, V., T. Sandholm, and J. Lang. 2007. When are elections with few candidates hard to manipulate? *Journal of the ACM* 54(3): 14.

Cramton, P., Y. Shoham, and R. Steinberg, ed. 2006. *Combinatorial auctions*. Cambridge, MA: MIT Press.

Feigenbaum, J., C. Papadimitriou, R. Sami, and S. Shenker. 2005. A BGP-based mechanism for lowest cost routing. *Distributed Computing* 18: 61–72.

Feigenbaum J., and S. Shenker. 2002. Distributed algorithmic mechanism design: Recent results and future directions. In *Proceedings of the international workshop on discrete algorithms and methods for mobile computing and communications*. 1–13.

Goldreich O., S. Micali and A. Wigderson. 1987. How to play any mental game or a completeness theorem for protocols with honest majority. In *Proceedings of the symposium on theory of computing*. 218–29.

Hajiaghayi, M.T., R. Kleinberg,and T. Sandholm. 2007. Automated online mechanism design and prophet inequalities. In *Proceedings of the national conference on artificial intelligence*.

Hershberger J, and S. Suri. 2001. Vickrey prices and shortest paths: What is an edge worth? In *Proceedings of the symposium on foundations of computer Science*. 252–9.

Jurca R, and B. Faltings. 2006. Minimum payments that reward honest reputation feedback. In *Proceedings of the ACM conference on electronic commerce*. 190–9.

Kfir-Dahav N, D. Monderer, and M. Tennenholtz. 2000. Mechanism design for resource bounded agents. In *Proceedings of the international conference on multi-agent systems*. 309–315.

Land, A., S. Powell, and R. Steinberg. 2006. PAUSE: A computationally tractable combinatorial auction. In *Combinatorial auctions*, ed. Cramton, Shoham, and Steinberg. Cambridge, MA: MIT Press.

Larson K. 2006. Reducing costly information acquisition in auctions. In *Proceedings of the autonomous agents and multi-agent systems*. 1167–74.

Larson K., and T. Sandholm. 2001. Costly valuation computation in auctions. In *Proceedings of the theoretical aspects of rationality and knowledge*. 169–182.

Larson K, and T. Sandholm. 2005. Mechanism design and deliberative agents. In *Proceedings of the international conference on autonomous agents and multi-agent systems*. 650–656.

Lavi R., and C. Swamy. 2005. Truthful and near-optimal mechanism design via linear programming. In *Proceedings of the symposium on foundations of computer science*. 595–604.

Lehmann D., R. Müller, and T. Sandholm. 2006. The winner determination problem. In P. Cramton, Y. Shoham and R. Steinberg.

Lehmann, D., L.I. O'Callaghan, and Y. Shoham. 2002. Truth revelation in rapid, approximately efficient combinatorial auctions. *Journal of the ACM* 49: 577–602.

Likhodedov A, and T. Sandholm. 2005. Approximating revenue-maximizing combinatorial auctions. In *Proceedings of the national conference on artificial intelligence*. 267–74.

Monderer D, and M. Tennenholtz. 1999. Distributed games: From mechanisms to protocols. In *Proceedings of the national conference on artificial intelligence*. 32–7.

Müller R. 2006. Tractable cases of the winner determination problem. In P. Cramton, Y. Shoham and R. Steinberg.

Nisan N, and A. Ronen. 2000. Computationally feasible VCG mechanisms. In *Proceedings of the ACM conference on electronic commerce*. 242–52.

Nisan, N., and A. Ronen. 2001. Algorithmic mechanism design. *Games and Economic Behavior* 35: 166–196.

Nisan, N., and I. Segal. 2006. The communication requirements of efficient allocations and supporting prices. *Journal of Economic Theory* 129: 192–224.

Parkes D., and J. Shneidman. 2004. Distributed implementations of generalized Vickrey–Clarke–Groves auctions. In *Proceedings of the international conference on autonomous agents and multi-agent systems*. 261–268.

Petcu A., B. Faltings, and D. Parkes. 2006. MDPOP: Faithful distributed implementation of efficient social choice problems. In *Proceedings of the international conference on autonomous agents and multi-agent systems*. 1397–404.

Rothkopf, M. 2007. Thirteen reasons why the Vickrey–Clarke–Groves process is not practical. *Operations Research* 55: 191–197.

Rothkopf, M., A. Pekěc, and R. Harstad. 1998. Computationally manageable combinatorial auctions. *Management Science* 44: 1131–1147.

Sandholm T. 1993. An implementation of the contract net protocol based on marginal cost calculations. In *Proceedings of the national conference on artificial intelligence*. 256–62.

Sandholm, T. 2000. Issues in computational Vickrey auctions. *International Journal of Electronic Commerce* 4: 107–129 .Early version in ICMAS-96.

Sandholm, T. 2002a. Algorithm for optimal winner determination in combinatorial auctions. *Artificial Intelligence* 135: 1–54 .Earlier versions: ICE-98 keynote, Washington U. tech report WUCS-99-01 Jan. 1999, IJCAI-99.

Sandholm, T. 2002b. eMediator: A next generation electronic commerce server. *Computational Intelligence* 18: 656–676 .Earlier versions: Washington U. tech report WU-CS-99-02 Jan. 1999, AAAI-99 Workshop on AI in Ecommerce, AGENTS-00.

Sandholm T. 2006. Optimal winner determination algorithms. In P. Cramton, Y. Shoham and R. Steinberg.

Sandholm, T. 2007. Expressive commerce and its application to sourcing: How we conducted $35 billion of generalized combinatorial auctions. *AI Magazine* 28(3): 45–58.

Sandholm T, and Boutilier C. 2006. Preference elicitation in combinatorial auctions. In P. Cramton, Y. Shoham and R. Steinberg.

Sandholm T, V. Conitzer, and C. Boutilier. 2007. Automated design of multistage mechanisms. In *Proceedings of the international joint conference on artificial intelligence*. 1500–6.

Sandholm T., and V. Ferrandon. 2000. Safe exchange planner. In Proceedings of the international conference on multi-agent systems. 255–62.

Sandholm, T., K. Larson, M. Andersson, O. Shehory, et al. 1999. Coalition structure generation with worst case guarantees. *Artificial Intelligence* 111: 209–238.

Shneidman J, and D.C Parkes . 2004. Specification faithfulness in networks with rational nodes. In *Proceedings of the ACM symposium on principles of distributed computing*. 88–97.

Vorobeychik Y, C. Kiekintveld, and M. Wellman. 2006. Empirical mechanism design: Methods, with application to a supply chain scenario. In *Proceedings of the ACM conference on electronic commerce*. 306–15.

Comte, Isidore Auguste Marie François Xavier (1798–1857)

Robert B. Ekelund Jr.

Auguste Comte, co-founder of sociology and positivist philosopher, was born at Montpellier in 1798 and died in Paris in 1857. A student at the Ecole Polytechnique until he was dismissed for disobedience and incorrigible behaviour, Comte became the secretary to Henri de Saint-Simon in 1818, a position he held until 1824. Over this period Comte developed the kernel of positivist philosophy and, along with Saint-Simon, modern sociology. The irascible Comte spent the rest of his life – often in the face of frequent poverty and desperate personal circumstances including mental breakdowns and a failed marriage – establishing, altering and working out positivism as a philosophical system of knowledge and as the foundation for the 'science of society'. Comte's *chefs d'oeuvre* included the encyclopedic *Cours de philosophie positive* (1830–42) and his *System of Positive Polity* (1851–4).

Comte's social philosophy encompassed all aspects of life, which he believed to be the harmonious working together of two inseparable elements – an organism and its environment. The 'true' methods of science were empirical (or inductive), and Comte, as did his mentor Saint-Simon, believed that they could be applied, *mutatis mutandis,* to all branches of thought including the social sciences, of which economics was a part. He succinctly described this basis in the *System of Positive Polity*: 'It [the positivist synthesis of all knowledge] rests at every point upon the unchangeable Order of the World. The right understanding of this order is the principal subject of our thoughts; its preponderating influence determines the general course of our feelings; its gradual improvement is the constant object of our actions . . .' (vol. 1, p. 21).

'Gradual improvement' came about in three successive stages of understanding the cause of phenomena. In the 'Law of Three Stages', Comte argued that man's conception of 'order' or the interrelationships of causes and effects, went first through a religious or theological stage, the theological–primitive content of which diminished progressively from fetishism, polytheism and monotheism. The metaphysical or *a priori* stage followed the theological and represented man's attempt to discover 'order' by reason. The final stage in understanding order was the positive stage wherein science, or the knowledge of relationships between disparate phenomena, brought man to a kind of perfection. Within this latter stage, Comte argued that a second great law obtained – that of decreasing generality and increasing complexity of understanding. Here, sciences progressed in a definite ordering, each dependent on the previous one, from mathematics to astronomy to physics to chemistry to biology and, finally, to sociology.

Comte's system, its broadly empirical methodology, and its precept that society was an organism which evolved under constraints that were themselves ultimately altered by social activities and behaviour, had a measurable impact upon classical economics and upon the *form* that

economic analysis would take in the neoclassical period and beyond. J.S. Mill introduced Comte's ideas to England and, for a time at least, was deeply influenced by the French philosopher. Mill's *Logic* (1843) contains copious and favourable references to Comte's inverse-deductive (empirical-historical) method for use in the broader and more important investigation of integrated social studies. In the end, however, Mill reaffirmed the *a priori* 'Ricardian' method for the narrower concerns of political economy. Again, Comte's influence surfaced in Mill's *Principles of Political Economy* (1848) in the form of the famous 'statics and dynamics' distinction and in Mill's admissions that the ultimate aim of social science was a broader conception of the process than political economy had to offer. In his *Principles,* however, Mill reaffirmed the (admittedly provisional) *a priori* deductive method as the essential and proper one for studying political economy. Ultimately, Mill and a number of Comte's followers totally rejected the 'religion of humanity' that Comte later made of his positivist principles (1851–77). Mill was especially aghast at the infringements on individual liberty that Comte's quasi-medieval 'Catholicism without Christianity' envisioned. In his *Autobiography* Mill attacked Comte's planned society as 'the completest system of spiritual and temporal despotism which ever yet emanated from a human brain, unless possibly that of Ignatius Loyola' (1873, p. 149).

In the late classical and early neoclassical period Comte's ideas received some reinforcement at the hands of the British historicists, notably John Kells Ingram. But his impact, especially on methodology and on his belief that political economy should be subordinate to sociology, was successfully nullified by the efforts of John Elliott Cairnes (1870). While admitting that empiricism was a necessary adjunct to economic theory, Cairnes defended an essentially abstract and *a priori* method in political economy. Cairne's ideas on method were replicated by the leading methodologist of the neoclassical and even post-neoclassical periods, John Neville Keynes. The attempt to infuse Comtian and other broader methods into political economy, in other words, reinforced Ricardo's method which became the dominant method of economists in the 20th century.

See Also

- ▶ Positivism
- ▶ Spencer, Herbert (1820–1903)

Selected Works

1830–42. *Cours de philosophie positive*, 2 vols, 5th ed. Trans. H. Martineau; 3rd ed. London, 1893.

1851–4. Système de politique positive, ou Traité de sociologie institutant la religion de l'humanite (Trans: J.H. Bridge et al. as *System of positive polity*). London, 1875–7; New York: Burt Franklin Reprint, 4 vols, 1968.

1899. *Lettres inédites de John Stuart Mill à Auguste Comte, publiées avec les réponses de Comte et un introduction par Lucien Lévy-Bruhl*. Paris: Felix Alcan.

References

Abrams, P. 1968. *The origins of British sociology 1834–1914*. Chicago: University of Chicago Press.

Adelman, P. 1971. Frederic Harrison and the 'Positivist' attack on orthodox political economy. *History of Political Economy* 3(1): 170–189.

Cairnes, J.E. 1870. M. Comte and political economy. *Fortnightly Review* 13: 579–602.

Ekelund Jr., R.B., and E. Olsen. 1973. Comte, Mill and Cairnes: The positivist-empiricist interlude in late classical economics. *Journal of Economic Issues* 7(3): 383–416.

Hayek, F.A. 1955. *The counter-revolution of science, studies in the abuse of reason*. New York: The Free Press.

Ingram, J.K. 1899. *A history of political economy*. London: A. and C. Black, 1915.

Mill, J.S. 1865. *Auguste Comte and positivism*. London: N. Trübner.

Mill, J.S. 1873. *Autobiography*. New York: Columbia University Press, 1924.

Seligman, B.B. 1969. The impact of positivism on economic thought. *History of Political Economy* 1(2): 256–278.

Concentration Measures

Juan Esteban Carranza

Abstract

Concentration is a characterization of the size distribution and quantity of competing firms within a specific market or industry. The most common concentration measures are the Herfindahl index and the *n*-firm concentration rate. The Herfindahl index is the sum of the squared market shares of all the firms in a market, whereas the *n*-firm concentration rate is the sum of the market shares of the *n* biggest firms. These measures are a significant reflection of the underlying degree of competitiveness, but are sensitive to the adopted market definition, and must be interpreted carefully depending on the specifics of the case.

Keywords

Gini coefficient; Herfindahl index; Lerner index; Market definition; Market structure

JEL Classification

D4

The term *concentration* (also *firm concentration*, *industry concentration* or *market concentration*) refers to aspects of the distribution of firm size within a specific market or industry that have traditionally been used to characterize the degree of competitiveness in the market. Even though the size of firms can be measured using many different variables, such as employment or assets, the sales level is the most commonly used size measure. Accordingly, if very few firms serve a very large portion of the market, it is said that the given market is highly 'concentrated', whereas if no single firm has a large share of sales it is said that the market is not 'concentrated'. Since concentration is an important reflection of the underlying market structure, its measurement is an important characterization of the interaction of firms within a specific market or industry.

The most common concentration measures are the '*n*-firm concentration rate' and the 'Herfindahl index'. Let S_i be the market share of firm i; the '*n*-firm concentration rate' is the sum of the market shares of the n biggest firms within the market:

$$C(n) = \sum_{i=1}^{n} S_i.$$

As indicated, the summation above is taken over the set of n biggest firms in the market. So, for example, the two-firm concentration rate of a given market is the sum of the market shares of the two biggest firms in the market where size is measured according to observed sales. In order to fully characterize the concentration of any given market, though, a number of these rates must be used, since there is no agreed-on value for n. This complicates its use for comparing concentration over time and across sectors, and for its use in statistical analysis.

The Herfindahl index, first devised by Albert Hirschman to measure the concentration of trade across sectors (so that the index is also known as 'Herfindahl–Hirschman index'; see Hirschman 1964, for its history), is the sum of the squared market shares of all firms in the market:

$$H = \sum_{i=1}^{N} S_i^2.$$

The summation in this case is taken over the set of all N firms in the market. This index lies between zero and 1: if there is only one firm in the market, so that the market has the highest possible concentration, the index is 1. If, on the other hand, there are many equally sized firms in the market, the index will be close to zero. By squaring the individual market shares, this index gives relatively greater weight to the market shares of large firms. Conversely, the addition of one small firm to the market dilutes somewhat the market share of larger firms, and has a marginal negative effect on the index, which is consistent with any notion of market concentration. Any value of this index can correspond to multiple

market configurations, being in that sense less illustrative of the actual concentration of a market than a set of *n*-firm concentration rates. On the other hand, this index can be easily correlated with other market characteristics and is therefore very useful for statistical analysis.

Other less commonly used concentration measures include entropy coefficients, the Gini coefficient and measures of the variance of market shares across firms within a market. The entropy coefficient is usually computed using the following formula:

$$E = \sum_{i=1}^{N} S_i log_2 \left(\frac{1}{S_i} \right).$$

This index takes value zero if there is only one firm in the market and grows as market concentration decreases. The interpretation of this coefficient is complicated, because its formula weights both large and small firms less heavily than mid-size firms and grows unboundedly as market concentration decreases. It is therefore less commonly used than the Herfindahl index.

The Gini coefficient is commonly used to characterize the income or wealth inequality within a society. Its drawback as a measure of market concentration is that it is useful only to measure the concentration of firms' sizes within a market, given a number of firms. So according to the Gini coefficient a duopolistic market with two firms of equal size is as concentrated as a market with 100 firms with identical size. The same drawback applies for the use of measures of the variance of firm size – whose definition is simply the sample variance of firm sizes.

All the concentration measures mentioned above are very sensitive to the actual market definition that is used. In markets for differentiated products, for example, products may face a continuum of similar products, and determining which similar products exactly constitute a market is not always easy. Take the specific example of the market faced by US mobile phone services: with just a handful of national providers it is concentrated given the standard concentration measures. These national firms, nevertheless, are also competing with local companies in various segments of the market and even with long-distance phone companies and Internet companies as providers of communication services. The Internet, on the other hand, competes in some instances with cable and satellite companies, radio stations and even newspapers as sources of news and entertainment. What exactly the relevant market faced by mobile phone companies is will depend on the type of issue being addressed. Accordingly, the concentration measures will change depending on the adopted market definition.

On the other hand, even if the market is well defined, computed concentration measures may not reflect at all the real competitive structure of the market. For example, even in markets as highly concentrated as the market for computer processors, the dominant firms have to account for the invisible competition of potential entrants. The same happens in regional markets where outside firms are kept at bay by few local firms with a combination of low prices and high transportation costs. Computed concentration measures for specific markets cannot account for this unobserved competition and may therefore lead to wrong conclusions regarding the underlying behaviour of firms. In these instances, a behavioural measure, such as the Lerner index, which measures the relative size of firms' markups, may be a better indicator of the competitive structure of the market.

There is a body of empirical literature that uses market concentration measures across industries to approximate the underlying differences in industries' competitiveness. They were then used to infer statistically the relationship between market 'structure' and market 'performance'. For example, correlations of R&D expenditure and market concentration were computed to investigate whether firms in concentrated markets were more or less likely to innovate than firms in more competitive markets. The value of such correlations is limited because the observed concentration may be both a cause and an effect of individual firms' behaviour and the relationship is shaped by the specifics of the industry. In order to avoid the ambiguities of such an inter-industry approach, the more recent empirical microeconomic literature has generally focused instead on the understanding of firm behaviour within specific industries, for which the use of concentration measures is less relevant.

See Also

- ▶ Competition
- ▶ Gini Ratio
- ▶ Market Structure

Bibliography

Hirschman, A.O. 1964. The paternity of an index. *American Economic Review* 54: 761–762.
Ravenscraft, D. 1983. Structure-profit relationships at the line of business and industry level. *Review of Economics and Statistics* 65: 22–31.
Scherer, F.M. 1970. *Industrial market structure and economic performance*, 2nd ed. Boston: Houghton Mifflin.
Simon, H., and C.P. Bonini. 1958. The size distribution of business firms. *American Economic Review* 48: 607–617.

Concentration Ratios

William G. Shepherd

These standard indicators of the degree of oligopoly in markets are used in studying market conditions. The ratio is the combined market shares of the 'oligopoly group' of firms which, being few, are closely interdependent. The ratio is usually based on three, four or five firms.

In theory the ratio indicates the market power held by the interdependent group. When the ratio is high, a few firms dominate the market and, with some degree of collusion, can raise prices and perform other conventional monopoly actions. Higher concentration makes effective collusion more probable. If the oligopolists achieve perfect joint maximation of profits, then the market power they exert is as great as if the firms were unified into one dominant firm. Concentration is therefore an indicator of diluted monopoly power.

How diluted it is depends on the degree of firms' cohesion. The fewer the firms, the more impact a departure from collusion will have on the joint outcome. Also, such departures will tend to be discovered more quickly, and therefore be open to more effective punishment by the other oligopolists. Therefore, higher concentration tilts the oligopolists' choices away from maverick price-cutting and toward collusion. Accordingly, high concentration should avoid the disintegration that often afflicts efforts to fix prices.

The distinction between tight oligopoly (a four-firm ratio above 60 per cent) and loose oligopoly (a ratio below 40 per cent) has come to be regarded as particularly important. In tight oligopoly, collusion is likely to crystallize effectively into strong cooperation, as the oligopolists' common interests overwhelm the rewards from cheating. Loose oligopoly, by contrast, is seen as a setting for disintegration, where the many oligopolists with low market shares are jointly unable to avert the endemic price cutting. This reasoning, which is broadly confirmed by the common run of experience in actual markets, suggests that a threshold value of concentration in the 50–60 per cent range should present a clear divide between the effective competition seen in loose oligopoly and the high market power that tight oligopoly may create.

Concentration is usually second in importance to individual market shares, as an element of market structure. Thus, for example, a General Motors may have a market share of 50 per cent, while it and the other three leading firms have a combined concentration ratio of 90 per cent. GM's own price–profit, innovation and other results are likely to be influenced more closely by its own 50 per cent share – giving it a dominant-firm position – than by the fact that it is in a market with high concentration in four firms.

Concentration's effects, such as they are, may be modified by entry conditions. Free and vigorous entry can limit the tight oligopolists' ability to exert market power. This is a controversial area, for there may be few active entrants actually ready to take advantage of free-entry conditions. Also, entry may be slow or marginal, so that it does not strongly affect the core of the market positions held by the leading firms.

Starting in the 1930s, the ratios soon gained pre-eminence in studying the degrees of market power. One was the then-new focus on oligopoly. The other was the new availability of the actual ratios for hundreds of US manufacturing industries for the year 1935. (US and UK ratios appear about every four or five years, as part of the

industrial census. Only manufacturing industries are covered, although ratios on urban US banking markets are prepared by other sources).

Econometric analysis has made extensive use of these ratios, with scores of papers reporting regressions relating concentration to price–cost patterns, growth rates, efficiency, rates of innovation, etc. In fact, concentration ratios reigned as the research focus for analysis of market power and its effects. Indeed, from 1939 to about 1970 the abundant availability of the ratios reinforced the tendency to regard (perhaps over-regard) oligopoly as a central issue. With the growing focus on individual market shares, the role of concentration ratios became less important in the 1970s. Even so, the ratios continue to be an indispensable descriptive statistic, used widely in research into the several elements of structure and their effects.

Before summarizing some of the results, one needs to note that the ratios are subject to a serious technical fault. A correct definition of the market is important, if true concentration within the market is to be measured. The US Bureau of the Census has a detailed system of standard industrial classification (SIC). All sectors are grouped by numbers, ranging from SIC 1 to 100. The manufacturing sector covers 'industry groups' 20 through 39, and so forth. Differing degrees of fineness are given. The five-digit product level is, on balance, about correct in fitting the average scope of true markets. Yet most research has focused on the much broader four-digit 'industries', which now number about 450 in the US.

About half of these four-digit census 'industries' depart seriously from correct market boundaries. The use of raw census ratios has undermined a good deal of the past research on structure and its effects. Commonly, the ratios are too broad, lumping together distinct products and geographic markets (for example, the national four-firm US ratio is 14 per cent for newspapers, but concentration ratios in true local newspaper markets probably average close to 100 per cent). Some cases go the other way, with the official ratios too high (for example, imports are not included in the ratios, and this is important for steel, television sets, shoes, cameras, automobiles and many others).

Adjusting the ratios to fit true market conditions requires care and judgement. Thus the weighted average degree of four-firm concentration in US manufacturing industries has been actually about 60 per cent, rather than the 40 per cent indicated by the raw census ratios.

Properly adjusted to eliminate these biases, a concentration ratio is an excellent descriptive statistic. It conveys the degree of oligopoly 'tightness'. It can show changes in structure pretty accurately. Thus the market power indicated by a ratio of 53 or 63 may be a matter of debate, but a rise in ratio from 53 to 63 for a given industry strongly suggests that there has been a rise in market power.

By 1975, the ratios have provided enormously valuable lessons, in several directions. They were the workhorse statistic in describing the structure of hundreds of industries, in case studies of industries, in antitrust investigations and in other straightforward treatments. They had become a pivotal basis for antitrust policy choices, such as in merger cases, where the degree of concentration was one basis for deciding whether to oppose the merger. During 1960–75 they gave rise to many scores of regression analyses, which tried to relate concentration to the possible effects of market power.

The correlation of concentration with price behaviour drew the largest volume of testing. These estimations were commonly plagued by the use of uncorrected, raw census ratios, which introduced substantial errors. Even so, a broad and significant correlation did emerge, enough to establish a presumption that the theoretical effects of oligopoly market power on pricing activity do occur in practice. The most successful testing involved multivariate models of five to eight independent variables, including filter variables such as capital intensity and growth, and other structure-related variables such as market shares, advertising intensity and capital-requirements barriers to entry.

Another important line of research attempted to explain concentration as the dependent variable. Growth rates were one possible causative variable, while economies of scale were another. Growth emerged as a very weak influence, and scale economies turned out to explain only a limited amount of actual concentration in important national markets. The weakness of these results may, in part, have reflected errors in the concentration ratios themselves.

Finally, concentration's possible effects on innovation, stability and wealth distribution

have been explored. These studies often were forced to adopt quite creative approaches, in light to the data problems. Again, the general patterns have confirmed the main predictions of theory, that high concentration tends to affect results much as high monopoly power does. High concentration was associated, on the whole, with slowed innovation, greater instability of production and the disequalizing of the distribution of wealth.

Yet testing continues on all of these points, and the concentration ratios have tended only to suggest patterns, not resolve them. Studies since 1970 have focused more on individual market shares, showing them to have stronger effects than concentration, just as theory predicts. Though the use of concentration ratios in regression analysis may have peaked, the ratios (adjusted as appropriate) will undoubtedly continue as a main basis for describing the degree of market power in a wide array of markets in the US, UK, Canada, Japan and certain other countries.

See Also

- ▶ Degree of Monopoly
- ▶ Market Share
- ▶ Market Structure

References

Bain, J.S. 1956. *Barriers to new competition*. Cambridge, MA: Harvard University Press.

Bain, J.S. 1968. *Industrial organization*, Revisedth ed. New York: Wiley.

Blair, J.M. 1972. *Economic concentration*. New York: Harcourt, Brace, Jovanovich.

Chamberlin, E.H. 1933. *The theory of monopolistic competition*, 8th ed. Cambridge, MA: Harvard University Press. 1962.

Collins, N., and L.E. Preston. 1968. *Concentration and price–cost margins in manufacturing industries*. Berkeley: University of California Press.

Comanor, W.S., and R.H. Smiley. 1975. Monopoly and the distribution of wealth. *Quarterly Journal of Economics* 89: 177–194.

Comanor, W.S., and T. Wilson. 1975. *Advertising and market power*. Cambridge, MA: Harvard University Press.

Goldschmid, H.J., H.M. Mann, and J.F. Weston (eds.). 1974. *Industrial concentration: The new learning*. Boston: Little, Brown.

Mann, H.M. 1966. Seller concentration, barriers to entry, and rates of return in thirty industries, 1950–1960. *Review of Economics and Statistics* 48: 296–307.

Mansfield, E. 1964. Industrial research and development expenditures. *Journal of Political Economy* 72: 319–340.

Scherer, F.M. 1980. *Industrial market structure and economic performance*, 2nd ed. Boston: Houghton Mifflin.

Shepherd, W.G. 1969. Market power and racial discrimination in white-collar employment. *Antitrust Bulletin* 14: 141–161.

Shepherd, W.G. 1970. *Market power and economic welfare*. New York: Random House.

Shepherd, W.G. 1979. *The economics of industrial organization*. Englewood Cliffs: Prentice-Hall.

Shepherd, W.G. 1982. Causes of increased competition in the US economy, 1939–1980. *Review of Economics and Statistics* 64: 613–626.

Stigler, G.J. (ed.). 1955. *Business concentration and price policy*. Princeton: Princeton University Press.

Stigler, G.J., and J.K. Kindahl. 1970. *The behavior of industrial prices*. New York: Columbia University Press for the National Bureau of Economic Research.

US Bureau of the Census. 1980. *Concentration ratios in manufacturing, 1977*, MC77 (SR) 2. Washington, DC: Government Printing Office.

Condillac, Etienne Bonnot de, Abbé de Mureau (1714–1780)

Peter Groenewegen

Keywords

Baudeau, N.; Condillac, E. B. de; Exchange; Galiani, F.; Le Trosne, G. F.; Physiocracy; Price; Turgot, A. R. J.; Value; Verri, P.

JEL Classifications

B31

Philosopher and economist. Born at Grenoble, the third son of a well-to-do aristocratic family, Condillac took his name from an estate purchased by his father in 1720. As a sickly child with poor eyesight he had little early education and was apparently still unable to read by the age of 12. After his father's death in 1727 he moved to Lyon to live with his oldest brother, continuing his education at its Jesuit college. Through this brother he may have first met Jean Jacques Rousseau, who was tutor to

his nephews in 1740 and became a life-long friend. His second brother, l'Abbé de Mably, took Condillac to Paris in *c*. 1733 to study theology at Saint Sulpice and the Sorbonne. He was ordained in 1740 and for the rest of his life 'ever faithful to the Christian church, would always wear his cassock, always remain l'Abbé' (Lefèvre 1966, p. 11).

For the next 15 years he lived the life of a Paris intellectual, studying the philosophy of Descartes, Malebranche, Leibniz and Spinoza, 'to whose speculative systems he formed a life-long aversion, preferring the English philosophers Locke (who particularly influenced his thinking), Berkeley, Newton and rather belatedly, Bacon' (Knight 1968, pp. 8–9). In this period he published the works which made his philosophical reputation: the *Essay on the Origin of Human Knowledge* (1746), the *Traité des Systèmes* (1749), his most famous philosophical work *Treatise on the Sensations* (1754) described as the 'most rigorous demonstration of the [18th-century] sensationalist psychology' (Knight 1968, p. 12) and his *Traité des Animaux* (1755).

Apart from giving him entry to the Paris salons, where at Mlle de Lespinasse's salon he is reputed to have first met Turgot, another life-long friend (Le Roy 1947, p. ix), his intellectual reputation gained him the position of tutor to Louis XV's grandson, the Duke of Parma. From 1758 to 1767 he resided in Parma. Because of its prime minister's economic development policies, inspired by a mixture of 'mercantilism, physiocracy and the ideas of Gournay', Condillac developed an interest in economic matters, an interest 'indirectly confirmed by his known contacts with the Italian political economists, Beccaria and Gherardo' (Knight 1968, pp. 231–2). In 1768 he returned to Paris, but by 1773 had retired to his estate of Flux near Beaugency, where he died in 1780. During the last decade of his life he published his *Cours d'Etudes* (1775), his work on economics (1776), a text on logic (1780) for use in Polish Palatinate schools, and commenced the unfinished *La Langue des Calculs* (1798). In 1752, he became a member of the Royal Prussian Academy; in 1768 after his return from Parma he was elected to the French Academy. His works have been frequently collected, most recently by Le Roy (1947–51).

The impetus for Condillac's writing *Le Commerce et le Gouvernement* has been ascribed to a desire to assist his friend Turgot in the difficulties he faced in 1775 as finance minister over the grain riots induced by his restoration of the free trade in grain (Le Roy 1947, p. xxv; Knight 1968, p. 232). This fits with the work's unqualified support for free trade in general and the grain trade in particular (1776, esp. pp. 344–5, which seems directly inspired by the Paris events of 1775). Writing the book may also be explained as a return favour for Turgot's assistance in getting Condillac (1775) published (cf. Knight 1968, pp. 13, 232). Despite Condillac's strong support for this major part of Physiocratic policy and his close adherence to other aspects of Physiocracy, his argument that manufacturing was productive brought critical replies from Baudeau and Le Trosne (1777). In this context it may be noted that his work bears little direct Physiocratic influence, the major influence being Cantillon (1755), the only work directly cited apart from Plumard de Dangeul (1754). It is, however, possible to detect some influence from the economics of Turgot, Galiani and Verri on the theory of value, price and competition (cf. Spengler 1968, p. 212).

As published, the work is divided into two parts. The first provides the elements of the science. Its starting point is the foundation of value, which Condillac finds in the usefulness of an object relative to subjective needs making relative scarcity the key variable determining value. Value is distinguished from price because price can only originate in exchange. It is determined by the competition between buyers and sellers guided by their subjective estimation of value. Gains from exchange arise from differences in value; for Condillac, value cannot exchange for equal value. Although Condillac did discuss the costs of acquiring commodities, his emphasis is on exchange, trade and price. Exchange presumes surplus production and a need for consumption. Hence trade inspires and animates production and is essential to increasing wealth. Only simple pictures of production are presented: farm labourers producing prime necessities of food and materials; artisans transforming raw materials into essentials and luxuries; traders who circulate these products at home and abroad. By this circulation trade distributes the annual

product and under competitive conditions settles its true prices. Condillac is more concerned with developing the institutions associated with trade: growth of towns and villages, money, banking, credit, interest and the foreign exchanges, the defence of property by government and hence the need for taxation, and the effects of restraints on trade, including the grain trade. The second part is almost completely devoted to examining effects of specific obstacles to trade ranging from war, tariffs, taxes, excessive government borrowing to luxury spending in the capital city and exclusive trading privileges. Moderate wants combined with complete freedom constitute his recipe for the best form of economic development.

Condillac's economic work received a mixed reception from later economists. J.B. Say (1805, p. xxxv) described it as an attempt 'to found a system of . . . a subject which [the author] did not understand'. Jevons (1871, p. xviii) praised Condillac's 'charming philosophic work [because] in the first few chapters . . . we meet perhaps the earliest distinct statement of the true connections between value and utility. . .'. Macleod (1896 described it as a 'remarkable work . . . utterly neglected but in scientific spirit . . . infinitely superior to Smith'. Since then, it has remained neglected even though as 'a good if somewhat sketchy treatise on economic theory and policy [it was] much above the common run of its contemporaries' (Schumpeter 1954, pp. 175–6).

Selected Works

1746. *An essay on the origin of human knowledge.* Trans. Thomas Nugent. London: Nourse, 1756.

1749. *Traité des Systèmes, où l'on en démèle les inconvénine et les advantages*. Paris/Amsterdam.

1754. *Treatise on the sensations.* Trans. B.S. Geraldine Czar. London: Favill Press, 1930.

1755. *Traité des Animaux, où après avoir fain des observations critique sur le sentiment de Descartes et sur celui de M. Buffon on entreprend d'expliquer leurs principales facultés . . .*. Amsterdam.

1755. *Cours d'Etude pour l'instruction du Prince de Parma, Aujourd'hui Ferdinand, Duc de Parma*. Parma (and Paris).

1776. *Le Commerce et le gouvernement considerés relativement l'un à l'autre.* In *Oeuvres complètes de Condillac*. Vol. 4. Paris: Brière, 1821.

1780. *The logic.* Trans. B.S. Joseph Neef. Philadelphia, 1809.

1798. *Le langue des Calculs, Ouvrage Posthume et élémentaire*. Paris.

Bibliography

Cantillon, R. 1755. *Essay on the nature of commerce in general*. Trans. B.S.H. Higgs. London: Macmillan, 1931.

Jevons, W.S. 1871. *Theory of political economy*. 4th ed. London: Macmillan, 1911.

Knight, I.F. 1968. *The geometric spirit. The Abbé de Condillac and the French enlightenment*. New Haven/London: Yale University Press.

Le Roy, G., ed. 1947–51. *Oeuvres philosophiques de Condillac.* Paris: Press Universitaires de France.

Le Trosne, G.F. 1777. *De L'intérêt Social, par rapport à la Valeur, à la circulation, à l'Industrie, & au commerce intérieur & extérieur: Ouvrage élémentaire dans lequel on discoute quelques Principes de M. l'Abbé de Condillac.* Paris.

Lefèvre, R. 1966. *Condillac ou la joie de vivre*. Paris: Editions Seghers.

Macleod, H.D. 1896. *The history of economics*. London: Bliss, Sands & Co..

Plumard de Danguel, L.J. 1754. *Remarques sur les avantages et les désavantages de la France et de la Gr. Bretagne par rapport au commerce et aux autres sources de la puissance des états.* Leyden (and Paris).

Say, J.-B. 1805. *A treatise on political economy of the production, distribution and consumption of wealth.* Trans. C.R. Prinsep, New American Edition. Philadelphia, 1880; reissued New York: Kelley, 1963.

Schumpeter, J.A. 1954. *History of economic analysis.* London: Allen & Unwin, 1959.

Spengler, J.J. 1968. Condillac, Etienne Bonnot de. In *Encyclopedia of the social sciences*, vol. 3, 2nd ed. Chicago: Chicago University Press.

Condorcet, Marie Jean Antoine Nicolas Caritat, Marquis de (1743–1794)

H. Moulin and H. Peyton Young

Keywords

Black, D.; Borda, J.-C. de; Condorcet, Marquis de; Condorcet's paradox; Impossibility

theorem; Independence of irrelevant alternatives; Maximum likelihood; Strategic voting; Voting rules

JEL Classifications
B31

Condorcet was a French mathematician and philosopher. With many of his fellow *encyclopédistes* he shared the conviction that social sciences are amenable to mathematical rigour. His pioneer work on elections, the *Essai sur l'application de l'analyse a la probabilité des decisions rendues a la pluralité des voix* (1785) is a major step in that direction.

The aim of the *Essai* is to 'inquire by mere reasoning, what degree of confidence the judgement of assemblies deserves, whether large or small, subject to a high or low plurality, split into several different bodies or gathered in one only, composed by men more or less wise' (*Discours préliminaire* to the *Essai*, p. iv).

In modern words, this is the jury problem: to decide whether the accused is guilty or not requires converting the opinions of several experts, with varying competence, into a single judgement. Systematic probabilistic computations for this problem occupy most of the *Essai*, often camouflaging the essential contributions. The opaqueness and technicality of the argument meant that a full recognition of its importance did not occur until more than 150 years later (Black 1958). Since then Condorcet's findings have strongly influenced modern social choice theorists (for example, Arrow, Guilbaud and Black), and still play a central role in many of its recent developments.

The starting point is that majority voting is the unambiguously best voting rule when only two candidates are on stage. This fact, whose modern formulation is known as May's theorem (May 1952) was clear enough to the encyclopedists, too. How, then, can we extend this rule to three candidates or more? The naive, yet widely used, answer is plurality voting (each voter casts a vote for one candidate; the candidate with most votes is elected). Both Condorcet and Borda (his colleague in the Academy of Sciences) raise the same objection against the plurality rule. Suppose, says Condorcet (*Discours préliminaire,* p. lviii) that 60 voters have the opinions shown in Table 1 about three candidates A, B, C.

Condorcet, Marie Jean Antoine Nicolas Caritat, Marquis de (1743–1794), Table 1

	23	19	16	2
Top	A	B	C	C
	C	C	B	A
Bottom	B	A	A	B

In the illustration, candidate A wins by plurality. Yet if we oppose A against *B* only, A loses (25 to 35) and in A against *C,* A loses again (23 to 37). Thus the plurality rule does not convey accurately the opinion of the majority. From these identical premises, Borda proposes his well-known scoring method (each candidate receives 2 points from a voter who ranks him first, 1 point from one who ranks him second, and none from one who ranks him last; hence *C* is elected with score 78), whereas Condorcet opens a quite different route.

Condorcet posits a simple binomial model of voter error: in every binary comparison, each voter has a probability $1/2 < P < 1$ of ordering the candidates correctly. All voters are assumed to be equally able, and there is no correlation between judgements on different pairs. Thus for Condorcet the relevant data is contained in the 'majority tournament' that results from taking all pairwise votes:

$$\text{B beats A, 35 to 25; C beats A, 37 to 23;}$$
$$\text{C beats B, 41 to 19.}$$

Condorcet proposes that the candidates be ranked according to 'the most probable combination of opinions' (*Essai,* p. 125). In modern statistical terminology this is a maximum likelihood criterion (see Young 1986).

In the above example the most probable combination is given by the ranking: CBA since the three statements C over B, C over A, B over A agree with the greatest total number of votes. Condorcet's ranking criterion implies that an

Condorcet, Marie Jean Antoine Nicolas Caritat, Marquis de (1743–1794), Table 2

23	17	2	10	8
A	B	B	C	C
B	C	A	A	B
C	A	C	B	A

alternative (such as C) that obtains a majority over every other alternative must be ranked first. Such an alternative, if one exists, is known as a 'Condorcet winner'.

As Condorcet points out, some configurations of opinions may not possess such a winner, because the majority tournament contains a cycle (a situation known as 'Condorcet's paradox'). He exhibits the example shown in Table 2.

Here A beats B, 33 to 27; B beats C, 42 to 18; C beats A, 35 to 25. According to Condorcet's maximum likelihood criterion, this cycle should be broken at its weakest link (A over B), which yields the ranking B over C over A. Therefore in this case B is declared the winner.

Somewhat later in the *Essai* (pp. 125–6), Condorcet suggests that one may compute the maximum likelihood ranking of n candidates by, first, choosing the $n(n-1)/2$ binary propositions that have the majority in their favour; then, if there are cycles, *successively deleting* those with smallest majorities until a complete ordering of the candidates is obtained. Unfortunately, for $n > 3$ this heuristic algorithm does not necessarily yield the ranking that accords with the greatest number of votes. An axiomatic characterization of Condorcet's rule is given in Young and Levenglick (1978).

Condorcet's idea of reducing individual opinions to all pairwise comparisons between alternatives proved essential to the aggregation of preferences approach initiated by Arrow (1951). The key axiom independence of irrelevant alternatives (IIA) requires that voting on a pair of candidates be enough to determine the collective opinion on this pair: this generalizes majority tournaments by dropping the symmetry across voters and across candidates. In this sense Arrow's impossibility theorem means that the Condorcet paradox is inevitable in any non-dictatorial voting method satisfying IIA.

Many more useful insights can be discovered in the *Essai*. For instance the issue of strategic manipulations, which has played a central role in the theory of elections since the late 1960s, is suggested in places, although it is never systematically analysed. For example, on page clxxix of the *Discours Preliminaire,* Condorcet criticizes Borda's method as more vulnerable to a '*cabale'*. His argument is supported by the modern game theoretical approach: whenever the configurations of individual opinions guarantee existence of a Condorcet winner, it defines a strategy-proof voting rule. This is one of the principal arguments in favour of Condorcet consistent voting rules, namely, rules electing the Condorcet winner whenever it exists (see, for example, Moulin 1983, ch. 4).

See Also

- Borda, Jean-Charles de (1733–1799)
- Social Choice
- Voting Paradoxes

Selected Works

1785. Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité voix. Paris.

Bibliography

Arrow, K. 1951. *Social choice and individual values*. New York: Wiley.

Black, D. 1958. *The theory of committees and elections*. Cambridge: Cambridge University Press.

May, K. 1952. A set of independent necessary and sufficient conditions for simple majority decision. *Econometrica* 20: 680–684.

Moulin, H. 1983. *The strategy of social choice*. Amsterdam: North-Holland.

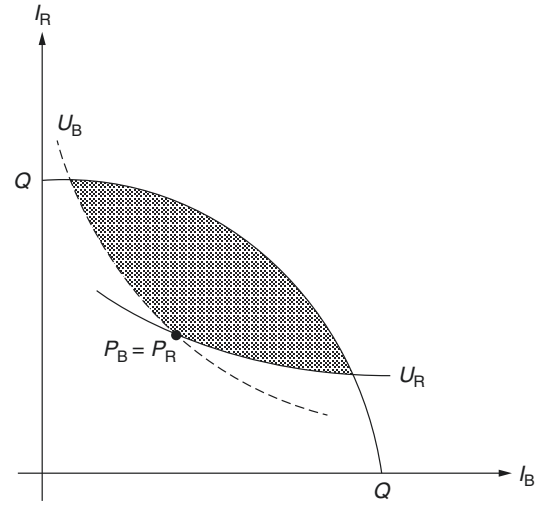
Young, H.P. 1986. Optimal ranking and choice from pairwise comparisons. In *Information pooling and group decisionmaking*, ed. B. Grofman and G. Owen. Greenwich: JAI Press.

Young, H.P., and A. Levenglick. 1978. A consistent extension of Condorcet's election principle. *SIAM Journal of Mathematics* 35: 285–300.

Conflict and Settlement

Jack Hirshleifer

All living beings are competitors for the means of existence. Competition takes the more intense form we call *conflict* when contenders seek to disable or destroy opponents, or even convert them into a supply of resources. Conflict need not always be violent; we speak, for example, of industrial conflicts (strikes and lockouts) and legal conflicts (law suits). But physical struggle is a relevant metaphor for these ordinarily non-violent contests.



Conflict and Settlement, Fig. 1 Statics of conflict – large potential settlement region

The Statics of Conflict

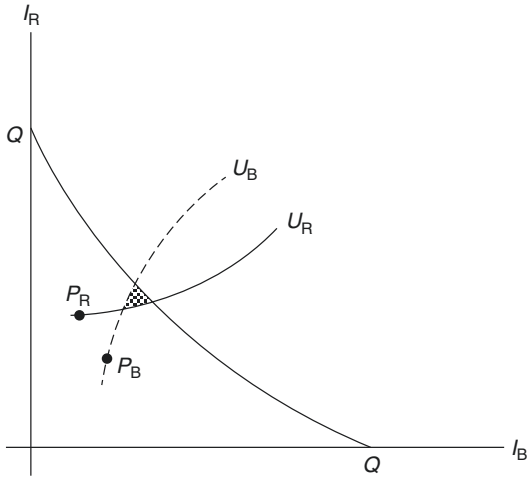
Involved in a rational decision to engage in conflict, economic reasoning suggests, will be the decision-maker's *preferences, opportunities* and *perceptions*. These three elements correspond to traditional issues debated by historians and political scientists about the 'causes of war': Is war mainly due to hatred and ingrained pugnacity (hostile preferences)? Or to the opportunities for material gain at the expense of weaker victims? Or is war mainly due to mistaken perceptions, on one or both sides, of the other's motives or capacities?

Of course it is quite a leap from the choices of individuals to the war-making decisions of collectivities like tribes or states. Group choice-making processes notoriously fail to satisfy the canons of rationality, most fundamentally owing to disparities among the interests of the individual members. Thus the internal decision-making structures of the interacting groups may also be implicated among the causes of war.

Setting aside this last complication, Figs. 1 and 2 are alternative illustrations of how preferences, opportunities, and perceptions might come together in a simple dyadic interaction. In each diagram the curve QQ bounds the 'settlement opportunity set' – what the parties can jointly attain by peaceful agreement or compromise – drawn on axes representing Blue's income I_B and Red's income I_R. The points P_B and P_R, in contrast, indicate the parties' separate *perceptions* of the income distribution resulting from conflict. The families of curves labelled U_B and U_R are the familiar utility indifference contours of the two agents.

Figure 1 shows a relatively benign situation: settlement opportunities are complementary, so there is a considerable mutual gain from avoiding conflict; the respective preferences display benevolence on each side; and the perceptions of returns from conflict are conservative and agreed (P_B and P_R coincide). The 'Potential Settlement Region' PSR (shaded area in the diagram), that is, the set of income distributions such that *both* parties regard themselves as doing better than by fighting, is therefore large – which plausibly implies a high probability of coming to an agreement. Figure 2 shows a less pleasant situation: antithetical opportunities, mutually malevolent preferences, and divergently optimistic estimates of the returns from conflict. The PSR is therefore small, and the prospects for settlement much poorer.

What might be called the *materialistic theory* attributes conflict, ultimately, to competition for resources. Primitive tribes attack one another for land, for hoards of consumables, or for slaves. Similar aims evidently motivated barbarian



Conflict and Settlement, Fig. 2 Statics of conflict – small potential settlement region

invasions of civilized cities and empires in ancient times, and European colonial imperialism in the modern era. Yet, between contending parties there will almost always be some element of complementary interests, an opportunity for mutual gain represented by the potential settlement region PSR. Orthodox economics has always emphasized the scope of mutual benefit, even to the point of losing sight of conflict; certain dissident schools, notably the Marxists, have committed the opposite error. While a detailed analysis cannot be provided here, among the factors underlying the relative material profitability of fighting versus negotiating are wealth differentials, Malthusian pressures, military technology, and the enforceability of agreements.

In contrast with the materialistic approach, *attitudinal theories* of conflict direct attention to the respective preference functions. An issue which has excited considerable interest concerns the relative weights assignable to genetic versus cultural determinants of attitudes. One extreme viewpoint, for example, regards xenophobic wars of family against family, of tribe against tribe, or nation against nation, as biologically 'normal' in the human species. An opposite interpretation pictures man as an innately compliant being, who has to be culturally indoctrinated into bellicosity.

Finally, what might be termed *informational theories* of conflict emphasize differences of perceptions of beliefs. Neoclassical economics tends to minimize the importance of such divergences – partly because they tend to cancel out from a large-numbers point of view, partly because incorrect beliefs are adjusted by experience in the process of establishing an economic equilibrium. But conflict and war are pre-eminently small-numbers, disequilibrium problems. Indeed, conflict may be regarded as in a sense an *educational process*. The school of actual struggle teaches the parties to readjust their perceptions to more realistic levels. Wars end by mutual consent when the potential settlement opportunities are seen as more attractive than continued fighting.

C

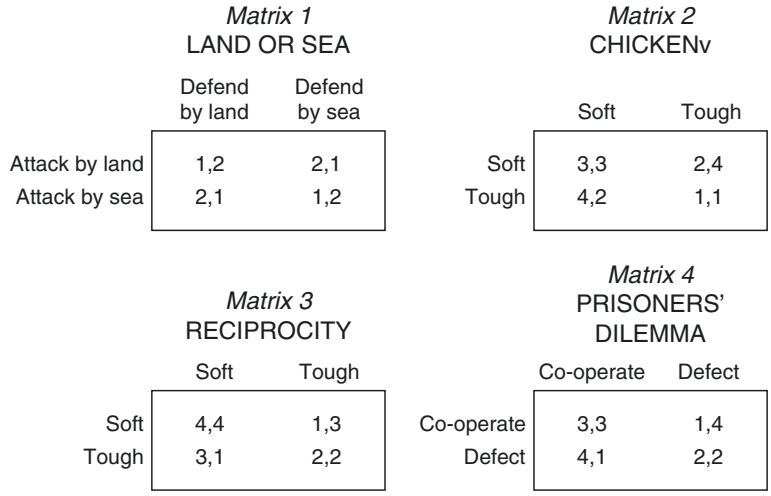
The Dynamics of Conflict

Static and dynamic elements are both importantly involved in conflict or settlement processes. In game theory terms, the *payoff environment*, represented by the familiar normal-form matrix, is the static element. The dynamic element may be called the *protocol of play*; as pictured in the game tree, the protocol specifies the allowable step-by-step moves in the light of the players' information at each stage.

A few very simple payoff environments are shown in Matrices 1–4. The numbers in each cell indicate ordinally ranked payoffs for each player, 1 being the poorest outcome in each case. In Matrix 1, 'Land or Sea', the environment is characterized by completely antithetical (constantsum) payoffs. The other three matrices – 'Chicken', 'Reciprocity' and 'Prisoners' Dilemma' – represent several of the many different possible mixed-motive situations combining an element of opposition of interests with an opportunity for mutual gain.

The simplest protocol to analyse is *one-round sequential play*: first Row selects one of his options, then Column makes his move in the light of Row's choice, and the game ends. In a sequential-play protocol it is always possible to find a 'rational' solution. If Column can be relied to choose his best final move then Row, knowing

Conflict and Settlement, Fig. 3



this, can calculate his best first move accordingly. (This process results in what is called a 'perfect equilibrium'.) In contrast, where the protocol dictates that players in a single-round game choose *simultaneously* – or, equivalently, where each chooses in ignorance of the other's move – solution concepts are harder to justify. The most commonly employed is called the 'Nash equilibrium' (or 'equilibrium point'), a pair of strategies from which neither player would want to diverge unilaterally.

In the 'Land or Sea' payoff environment, under the one-round *sequential-move* protocol, it is the second-mover or defender who has the advantage. If Row moves first, for example, Column can always successfully counter; e.g. if Row attacks by land, Column will defend by land. Hence the (1,2) payoff-pair is the outcome regardless of Row's initial move. In military terms the defence has an intrinsic advantage whenever the attacker must visibly commit his forces to one or another line of attack. And, of course, where the defence has such an advantage neither party is motivated to initiate warfare through aggression. But if 'Land or Sea' is played under the *simultaneous-move* protocol, both parties are groping in the dark and little can be said with confidence. (Here the Nash equilibrium would have each side choosing its move at random, in effect tossing a coin.)

In the payoff environment of 'Chicken' (Matrix 2), while the opportunities remain highly antithetical there is now a mutual interest in avoiding the disastrous (1,1) outcome that comes about when both play Tough. In contrast with 'Land or Sea', in the 'Chicken' payoff environment the advantage lies with the *first-mover*. Specifically, Row should rationally play Tough, knowing that Column then has to respond with Soft. For, Column must accept the bad (payoff of 2) to avoid the worst (payoff of 1). If the protocol dictates simultaneous moves, however, once again the players are groping in the dark. Under the Nash equilibrium concept they choose probabilistically, which implies that the disastrous (1,1) outcome will indeed occur a percentage of the time. There is a suggestive application of this model to industrial conflict. If union (or management) becomes committed to play Soft, it will be at a disadvantage in negotiations – the other side will then surely play Tough. But if both play Tough, there is no hope for peaceful settlement. Hence each side should rationally adopt a 'mixed' strategy, with the consequence that strikes and lockouts will occur in a certain fraction of the dealings.

The 'Reciprocity' payoff environment (Matrix 3) is more rewarding to cooperative behaviour. The idea is that each player would

Conflict and Settlement, Fig. 4

Matrix 5
DETERRENCE WITHOUT COMMITMENT

	Fold	Retaliate
Refrain	2,3	2,3
Attack	3,1	1,2

Matrix 6
DETERRENCE REQUIRING COMMITMENT

	Fold	Retaliate
Refrain	2,3	2,3
Attack	3,2	1,1

answer Soft with Soft – leading to the mutually preferred (4,4) payoffs – but failing this, would respond to Tough with Tough. If the *sequential-move* protocol applies, the first-mover would then always rationally choose Soft, and so the ideal (4,4) payoff-pair should be achieved. But under the *simultaneous-move* protocol, with each party in the dark about the other's move, again the outcome is quite unclear. In fact there are three Nash equilibria: pure-strategy solutions at (4,4) and (2,2), and a mixed-strategy solution as well.

Finally, in the famous 'Prisoners' Dilemma' payoff environment (Matrix 4) the parties are likely to find themselves in the Defect-Defect 'trap' with (2,2) payoffs, even though (3,3) could be achieved were each to play Cooperate. Here the 'trap' takes hold under both sequential-move and simultaneous-move protocols.

The preceding discussion could only be suggestive, limited as it was to 2-player single-round games, within that category to only a few two-strategy symmetrical payoff environments, and finally to the very simplest protocols – excluding, for example, all negotiations and communications between the parties. Space limitations permit comment upon only a few additional points:

Perceptions

Standard game models assume that players know not only their own payoffs but also their opponents'. Unintentional error on this score, or else deliberate deception, may play a crucial role. Suppose two parties in the 'Reciprocity' payoff environment of Matrix 3 find themselves initially playing Tough–Tough with outcome (2,2). Imagine now they are given a chance to shift strategies under a *sequential-move* protocol. As first-mover, Row would be happy to change from Tough to Soft if only he could rely upon Column to respond in kind. But Row may, mistakenly, believe that Column's payoffs are as in 'Chicken', from which he infers that Column would stand pat with Tough. Row would therefore not shift from Tough, hence Column in his turn would not change either. (Some authors have gone so far as to attribute all or almost all of human conflict to such mistaken 'self-fulfilling beliefs' about the hostility of opponents, but of course this pattern is only one of many possibilities.)

Commitment and Deterrence

In some circumstances the second-mover in point of time (Column) may be able to *commit* himself to a given response strategy before Row makes his first move. While Column thereby surrenders freedom of choice, doing so may be advantageous. Consider threats and promises. A *threat* is a commitment to undertake a second-move punishment strategy even where execution thereof is costly. A *promise* similarly involves commitment to a costly reward strategy. Matrices 5 and 6 illustrate how a threat works. Row's choices are Attack or Refrain, while Column's only options are to Retaliate or Fold if Row attacks. Column's problem, of course, is to deter Row's attack. In Matrix 5 Column prefers to Retaliate if attacked, a fact that – given Row's preference – suffices for deterrence. Commitment is not required. (Since Column prefers to Retaliate, there is no need to *commit* himself to do so.) In Matrix 6 the Column player prefers to turn the other cheek; if attacked,

he would rather Fold than Retaliate. Unfortunately, this guarantees he will be attacked! (Note that here it is not excessive hostility, but the reverse, that brings on conflict.) But if Column could *commit* himself to Retaliate, for example by computerizing the associated machinery beyond the possibility of his later reneging, then deterrence succeeds. In short, if a pacific player can reliably *threaten* to do what he does not really want to do, he won't have to do it! (Needless to say, so dangerous an arrangement is not to be casually recommended.)

The Technology of Struggle

Conflict is a kind of 'industry' in which different 'firms' compete by attempting to disable opponents. Just as the economist without being a manager or engineer can apply certain broad principles to the processes of industrial production, so, without claiming to replace the military commander, he can say something about the principles governing how desired results are 'produced' through violence.

Battles typically proceed to a definitive outcome – victory or defeat. *Wars* on the whole tend to be less conclusive, often ending in a compromise settlement. These historical generalizations reflect the working of increasing versus decreasing returns applied to the production of violence:

1. Within a sufficiently small geographical region such as a battlefield, there is a critical range of increasing returns to military strength – a small increment of force can make the difference between victory and defeat.
2. But there are decreasing returns in projecting military power away from one's base area, so that it is difficult to achieve superiority over an enemy's entire national territory. The increasing-returns aspect explains why there is a 'natural monopoly' of military force *within* the nation-state. The diminishing-returns aspect explains why a multiplicity of nation-states have remained militarily viable to this date. (However, there is some reason to believe, the technology of attack through long-range weapons has now so come to prevail over the defence that a single world-state is indeed impending.)

Going into the basis for increasing returns, at any moment the stronger in battle can inflict a more-than-proportionate loss upon his opponent, thus becoming progressively stronger still. Important special cases of this process are modelled via Lanchester's equations. In combat, in the ideal case where all the military units distribute their fire equally over the enemy's line, the process equations are:

$$\mathrm{d}\boldsymbol{B}/\mathrm{d}t = -k_{\mathrm{R}}\mathbf{R}$$
$$\mathrm{d}\boldsymbol{R}/\mathrm{d}t = -k_{\mathrm{B}}\mathbf{B}$$

Here B and R are the given force sizes for Blue and Red, and the per-unit military efficiencies are given by the k_{B} and k_{R} coefficients. It follows that military strengths are equal when:

$$k_{\mathrm{B}}\mathrm{B}^2 = -k_{\mathrm{R}}\mathrm{R}^2$$

But even where military strength varies sensitively than as the square of force size, it remains quite generally the case that in the combat process the strong become stronger and the weak weaker, leading to ultimate annihilation unless flight or surrender intervene. (Of course, a skilful commander finding himself with an adverse force balance will attempt to change the tactical situation – by timely withdrawal, deception, or other manoeuvre.)

One implication of increasing returns may be called the 'last-push principle'. In the course of a conflict each side will typically not be fully aware of the force size and strength that the opponent is ultimately able and willing to put in the field. Hence the incentive to stand fast, even at high cost, lest a potentially won battle be lost. (Foch: 'A battle won is a battle in which one will not confess oneself beaten'.) This valid point unfortunately tends to lead to battlefield carnage

beyond all reasonable prior calculations, as experienced for example at Verdun.

On the other hand, an effective substitute for force size is superior *organization*. An integrated military unit is far more powerful than an equally numerous conglomeration of individual fighters, however brave. Organizational superiority, far more than superiority in weapons, explains why small European expeditionary contingents in early modern times were able to defeat even vast indigenous forces in America, Africa and Asia. Battles are thus often a contest of organizational forms; the army whose command structure first cracks under pressure is the loser.

As for diminishing returns, in the simplest case an equilibrium is achieved at a geographical boundary such that:

$$M_B - s_B x_B = M_R - s_R x_R$$

Here M_B and M_R are military strengths at the respective home bases, S_B and S_R are decay gradients, and x_B and x_R are the respective distances from base. The condition of equality determines the allocation of territory.

The 'social physics' of struggle is of course far more complex than these simplistic initial models suggest. There are more or less distinct offence and defence technologies, first-strike capability is not the same as retaliatory strength, countering insurgency is a different problem from central land battle, etc.

Conflict, Society and Economy

Conflict theory can help explain not only the size and shape of nations, but the outcomes of competition in all aspects of life: contests among social classes, among political factions and ideologies, between management and labour, among contenders for licences and privileges ('rent-seeking'), between plaintiffs and defendants in law suits, among members of cartels like OPEC, between husband and wife and sibling and sibling within the family, and so on. Whenever resources can be seized by aggression, invasion attempts can be expected to occur. Invasive and counter-invasive effort absorb a very substantial fraction of society's resources in every possible social structure, whether egalitarian or hierarchical, liberal or totalitarian, centralized or decentralized. Furthermore, every form of human social organization, whatever else can be said for or against it, must ultimately meet the survival test of internal and external conflict.

Notes on the Literature of Conflict (of Special Relevance for Economists)

Classical military thought from Machiavelli to Clausewitz to Liddell Hart, though rarely analytical in the economist's sense, remains well worth study. An excellent survey is Edward Mead Earle (1941). Modern work in this classical genre understandably concentrates upon the overwhelming fact of nuclear weaponry and the problem of deterrence; the contributions of Herman Kahn (1960, 1962) are notable. There is of course a huge historical literature on conflict and war. An interesting economics-oriented interpretive history of modern warfare is Geoffrey Blainey (1973). William H. McNeill (1982) examines the course of military organization and technology from antiquity to the present, emphasizing the social and economic context. On a smaller scale John Keegan (1976) provides a valuable picture of how men, weapons, and tactics compete with and complement one another on the battlefield. There is also a substantial body of statistical work attempting in a variety of ways to summarize and classify the sources and outcomes of wars; the best known is Lewis F. Richardson (1960b). Mathematical analysis of military activity, that is, quantifiable modelling of the clash of contending forces, is surprisingly sparse. The classic work is Frederick William Lanchester (1916 [1956]).

The modern analysis of conflict, typically combining the theory of games with the rational-decision economics of choice, is represented by three important books by economists: Thomas C. Schelling (1960), Kenneth E. Boulding

(1962), and Gordon Tullock (1974). Works by non-economists that are similar in spirit include Glenn H. Snyder and Paul Diesing (1977) and Bruce Bueno de Mesquita (1981). A tangentially related literature, making use of the rather mechanical psychologistic approach of Richardson (1960a), includes a very readable book by Anatol Rapoport (1960).

See Also

- ▶ Bargaining
- ▶ Game Theory

References

Blainey, G. 1973. *The causes of war*. New York: The Free Press.

Boulding, K.E. 1962. *Conflict and defense: A general theory*. New York: Harper & Brothers.

Bueno de Mesquita, B. 1981. *The war trap*. New Haven: Yale University Press.

Earle, E.M. (ed.). 1941. *Makers of modern strategy: Military thought from Machiavelli to Hitler*. Princeton: Princeton University Press.

Kahn, H. 1960. *On thermonuclear war*. Princeton: Princeton University Press.

Kahn, H. 1962. *Thinking about the unthinkable*. New York: Avon Books.

Keegan, J. 1976. *The face of battle*. New York: Viking.

Lanchester, F.W. 1916. Aircraft in warfare: The dawn of the fourth arm. London: Constable. Extract reprinted in the world of mathematics, ed. James R. Newman, vol. 4. New York: Simon & Schuster, 1956, 2138–2157.

McNeill, W.H. 1982. *The pursuit of power: Technology, armed force, and society since AD 1000*. Chicago: University of Chicago Press.

Rapoport, A. 1960. *Fights, games, and debates*. Ann Arbor: University of Michigan Press.

Richardson, L.F. 1960a. *Arms and insecurity: A mathematical study of the causes and origins of war*. Pittsburgh: Quadrangle.

Richardson, L.F. 1960b. *Statistics of deadly quarrels*. Pittsburgh: Quadrangle.

Schelling, T.C. 1960. *The strategy of conflict*. Cambridge, MA: Harvard University Press.

Snyder, G.H., and P. Diesing. 1977. *Conflict among nations: Bargaining, decision making, and system structure in international crises*. Princeton: Princeton University Press.

Tullock, G. 1974. *The social dilemma: The economics of war and revolution*. Blacksburg: University Publications.

Congestion

Richard Arnott and Marvin Kraus

Abstract

'Congestion' is the phenomenon whereby the quality of service provided by a *congestible facility* degrades as its aggregate usage increases, when its capacity is held fixed. Here, the economic theory of congestion is developed in the context of road traffic. The primary questions of interest are how the capacity of a congestible facility and its usage fee should be chosen. This leads naturally to the question of whether the usage fees collected will be sufficient to cover capacity costs at the optimum.

Keywords

Clubs; Congestion; Externality cost; First-best pricing; Local public goods; Ramsey pricing; Second-best theory

JEL Classifications

R41

'Congestion' is the phenomenon whereby the quality of service provided by a *congestible facility* degrades as its aggregate usage increases, when its capacity is held fixed. We shall develop the economic theory of congestion in the context of road traffic, but congestion is pervasive: more telephone usage increases the probability of encountering a busy line; higher electricity demand may lead to voltage fluctuations, brownouts and eventually blackouts; more swimmers in a pool make comfortable swimming more difficult; more patients visiting a medical clinic results in longer waits and lower-quality care; in a more crowded classroom, students receive less individual attention, and more time is wasted on administration and discipline; and so on. The economic theory of congestion identifies how the capacity of a congestible facility and its usage fee should be

chosen. Some degree of congestion is typically socially optimal.

The economic theory of congestion has much in common with the theory of clubs and local public goods (Scotchmer 2002). The two literatures examine similar issues, but the economic theory of congestion has a policy perspective, while the theory of clubs and local public goods focuses on decentralized provision.

Formally, we may define congestion as follows. Consider a congestible facility in a steady state, that comprises I congestible elements. (Congestible elements for a sports stadium, for example, include nearby roads, parking facilities, the ticket office, washrooms, concessions, and seating.) Element i is characterized by a flow capacity, k_i, and a stock capacity, K_i, the flow capacity is the maximum throughput per unit time, the stock capacity the maximum number of users at a point in time. Similarly, the level of usage is described in terms of the throughput of congestible element i, n_i, and the number of users at a point in time, N_i. The congestible facility provides J dimensions of quality of service, with the level of dimension j indicated by sj.

Letting k, K,n, N and s denote the corresponding vectors,

$$s = S(k, K, n, N). \tag{1}$$

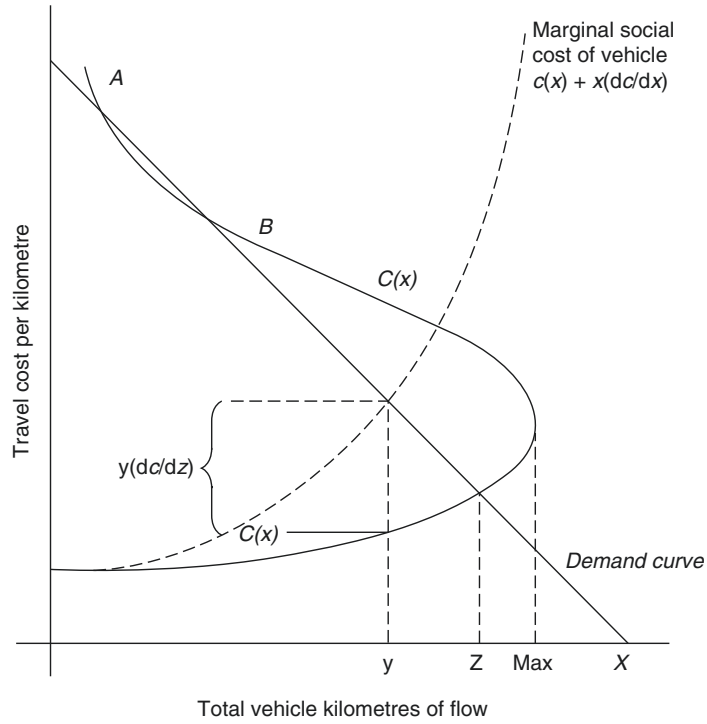
Congestion occurs when there is at least one combination of j and i for which s_j is monotone decreasing in n_i (flow congestion) or N_i (stock congestion), that is, when some dimension of quality of service falls as the throughput or stock of users of some congestible element of capacity increases. This is the static or steady-state definition of congestion. The dynamic definition of congestion adds time subscripts to s, k, K, n and N, and appends equations of motion relating stocks and flows for the various elements of capacity.

For some congestible elements, such as a turnstile, the bottleneck in the Vickrey (1969) bottleneck model of traffic congestion, or a switching circuit, the flow capacity constraint is the more important; for others, such as a telephone line, an elevator, a swimming pool, or seating at a football stadium, the stock capacity constraint is the more important. It should also be noted that a congestible facility can take the form of a network of congestible elements of capacity; a natural distinction is then between link congestion (for example, highway links) and nodal congestion (for example, traffic intersections).

To develop the theory, we consider a particular congestible facility having a single element of capacity and identical users, that is in a steady state: a road of uniform width connecting a single entry point A and a single exit point B, for which an increase in traffic flow increases travel time and an increase in road width reduces it. In this context, the deterioration of quality of service with an increase in usage is the increase in travel time from an increase in traffic flow.

We start with the short-run problem of determining optimal flow and its decentralized attainment, holding road width fixed. Let f denote flow, w road width, $t = t(f, w)$ the travel time function with (functional subscripts denote partial derivatives) $t_f > 0$ and $t_w < 0$ and p the value of time. Then the cost to an individual driver of travelling from A to B, the *user cost,* is $\rho\, t(f, w)$. Total user costs per unit time equal flow times user cost: $\rho f t(f, w)$. The social cost per unit time from increasing flow by one unit, with capacity held fixed, the *short-run marginal social cost,* is $\rho t(f, w) + \rho f t_f(f, w)$. The first term is the user cost of the extra driver; the second, the *congestion externality cost*. A driver imposes a congestion externality by slowing other drivers down; increasing steady-state flow by one car increases each car's travel time by $t_f(f, w)$ and social cost by $p f t_f(f, w)$.

Figure 1 displays short-run equilibrium. p denotes trip price, $D(p)$ the aggregate trip demand function, and $uc(f)$ and $srmsc(f)$ the user cost and short-run marginal social cost as a function of f, holding w fixed. With no toll, a user's trip price equals his user cost, and equilibrium occurs where the demand and user cost functions intersect, with flow f^e. Assuming that the marginal social benefit from a trip equals the corresponding marginal willingness to pay, the optimum occurs where the demand and short-run marginal social cost curves intersect, with flow f^*. Thus, with no toll, equilibrium flow is excessive.



Congestion, Fig. 1 Demand and supply on a congested highway

Efficiency obtains when economic agents face the social costs of their decisions and derive the social benefits from them. In the notoll case, the price of a trip falls short of its marginal social cost since a driver does not pay for slowing down other drivers. Following Pigou (1947), the standard remedy for internalizing the congestion externality is to impose a toll equal to the congestion externality cost, evaluated at the social optimum: τ^* in Fig. 1. This causes the trip price function to shift up from $uc(f)$ to $uc(f) + \tau^*$ and equilibrium flow to fall to the optimal level.

The above argument illustrates the general principle that efficient utilization of a congestible facility requires that the price equal short-run marginal social cost and the toll the congestion externality cost. Different user types - for example, cars and trucks – may impose different congestion externality costs. Efficiency then requires that the toll be differentiated according to user type.

We now turn to the long-run planning problem in which both road width and flow are choice variables. We then consider decentralization of the optimum. Let $B(f)$ denote the social benefit per unit time from flow f, and $C(w)$ the amortized capital cost of road width w. (We ignore the complications that arise when the congestible facility is sufficiently large that its construction alters factor prices.) The social surplus generated by the road (per unit time) equals social benefit minus social cost, and social cost equals total user cost plus amortized capital cost:

$$SS(f,w) = B(f) - \rho ft(f,w) - C(w). \quad (2)$$

It is easily seen from (2) that the road width that maximizes social surplus is that which minimizes social cost. This means that, when the long-run planning problem is solved, production is carried out according to the long-run cost structure, and the short-run marginal cost pricing (which is again) required for optimal flow is equivalent to long-run marginal cost pricing:

$$p = LRMC. \quad (3)$$

Now, recall the basic result of production theory that $LRMC$ is equal to, less than or greater than $LRAC$ (long-run average cost) according to

whether *LRAC* is constant, decreasing or increasing. Combining this with (3), we have the result that, when *LRAC* is constant, $p = LRAC$ holds at a long-run optimum. This is equivalent to equality between the total value of output and the total cost of output. Since total user cost is a component of both, this equality implies equality between toll receipts and amortized capital cost. Thus, *in the case of constant long-run average cost, the revenue raised from the optimal toll exactly covers the capital cost of providing a road of optimal width.* This is known as the 'self-financing' result. It was first derived by Mohring and Harwitz (1962) and subsequently generalized by Strotz (1965) (For a geometric derivation, see Arnott and Kraus 2003).

The self-financing result extends to congestible facilities with multiple elements of capacity, multiple dimensions of quality of service, and multiple user groups. If a congestible facility exhibits constant long-run average costs, provision of the facility can be decentralized via competing 'clubs'; competition will result in each club charging each user a fee for use of its congestible facility equal to the congestion externality cost he imposes, and choosing optimal capacity.

The above theory was developed on the assumption of a steady state. In the extension to treat nonstationary dynamics, which is conceptually straightforward, the distinction between flow externalities and stock externalities becomes sharper.

The theory relates to *first-best* pricing and capacity choice when congestion is the only externality. When usage entails other externalities, such as pollution, firstbest pricing should take these into account. In any policy context, additional practical constraints that rule out attainment of the full first-best allocation need to be considered. These are treated by applying second-best theory (Diamond and Mirrlees 1971). Consider, for example, the pricing problem facing a public transit authority. The underpricing of urban auto travel may call for the underpricing of mass transit (Lévy-Lambert 1968; Marchand 1968); since optimal lump-sum redistribution is infeasible for informational reasons, the authority may choose to sacrifice some efficiency to improve equity by charging lower fares to needy groups (Atkinson and Stiglitz 1980), rationing, or nonlinear pricing (Wilson 1993); administrative costs may preclude fine-tuning the fare according to distance travelled or time of day, leading to variants of Ramsey pricing (Mohring 1970); the authority may face a deficit constraint, requiring it to price above marginal social cost (Boiteux 1956); with distortionary taxation, the social cost of financing an extra dollar of transit authority deficit may significantly exceed one dollar (Vickrey 1959); and the government may choose to deviate from marginal social cost pricing to provide the public transit authority with higher-powered incentives (Laffont and Tirole 1993) or to achieve political objectives. These considerations will also cause second-best capacity to deviate from first-best capacity.

See Also

- ▶ Consumption Externalities
- ▶ France, Economics in (After 1870)
- ▶ Network Goods (Empirical Studies)
- ▶ Network Goods (Theory)
- ▶ Pigouvian Taxes
- ▶ Urban Economics
- ▶ Value of Time

Bibliography

Arnott, R., and M. Kraus. 2003. Principles of transport economics. In *Handbook of transportation science*, ed. R. Hall, 2nd ed. Boston: Kluwer.

Atkinson, A., and J. Stiglitz. 1980. *Lectures on public economics*. New York: McGraw-Hill.

Boiteux, M. 1956. Sur la gestion des monopoles publics astreints à l'équilibre budgétaire. *Econometrica* 24: 22–40.

Diamond, P., and J. Mirrlees. 1971. Optimal taxation and public production I: Production efficiency and II: Tax rules. *American Economic Review* 61 (8–27): 261–278.

Laffont, J.-J., and J. Tirole. 1993. *A theory of incentives in procurement and regulation*. Cambridge, MA: MIT Press.

Lévy-Lambert, H. 1968. Tarification des services à qualité variable: Application aux péages de circulation. *Econometrica* 36: 564–574.

Marchand, M. 1968. A note on optimal tolls in an imperfect environment. *Econometrica* 36: 575–581.

Mohring, H. 1970. The peak-load problem with increasing returns and pricing constraints. *American Economic Review* 60: 693–705.
Mohring, H., and M. Harwitz. 1962. *Highway benefits: An analytical framework*. Evanston: Northwestern University Press.
Pigou, A. 1947. *A study in public finance*. 3rd ed. London: Macmillan.
Scotchmer, S. 2002. Local public goods and clubs. In *Handbook of public economics*, ed. A. Auerbach and M. Feldstein, vol. 4. Amsterdam: North-Holland.
Strotz, R. 1965. Urban transportation parables. In *The public economy of urban communities*, ed. J. Margolis. Washington, DC: Resources for the Future.
Vickrey, W. 1959. Statement on the pricing of urban street use. *Hearings*: *US Congress*, *Joint Committee on Metropolitan Washington Problems*, 11 Nov, 454–477.
Vickrey, W. 1969. Congestion theory and transport investment. *American Economic Review Proceedings* 59: 251–260.
Wilson, R. 1993. *Nonlinear pricing*. Oxford: Oxford University Press.

Conglomerates

Alan Hughes

The overall output of a firm may be composed of activity in more than one product market. The growth of individual firms will be composed of changes in the scale of their activities in each of the markets in which they operate and in the numbers of those markets. In any period these changes will consist of horizontal expansion in the market(s) in which they operated at the beginning of the period and entry into new markets; where there is a supplier or buyer relationship with the original market then this expansion will be vertical integration. Expansion which fits neither of these categories is termed diversifying or conglomerate expansion. Growth in any of these directions may be by the purchase of new assets (internal growth), or by the purchase of existing assets through takeover or merger (external growth). Although it is common to refer to non-horizontal and non-vertical expansion as diversified, or conglomerate, the latter term also has a more specific connotation emphasizing particularly diverse external expansion. It has in particular been used to mean a company which has by a deliberate strategy of external growth, often away from declining sectors, developed a highly diversified product range which cannot easily be characterized in terms of a single, or well defined, group of production technologies, a single set of major competitors, or a stable place in a well defined industry group (Steiner 1975; Weston 1980). (And in the US context to have financed that expansion with issues of paper rather than cash, accompanied by accounting techniques for consolidating acquired companies designed to boost earnings per share and make future paper issues even more profitable (Blair 1972; Steiner 1975).)

Although diversified businesses predate World War II (often in the form of financial holding companies in Europe and Japan) conglomerate companies in the above sense are essentially a postwar phenomenon, and have been associated with the widespread adoption of decentralized divisionalized management structures. In particular, in the United States the growth of merger activity in the 1960s was dominated by diversified acquisitions, the most spectacular of which were associated with the emergence of a group of particularly aggressive conglomerates. For instance, between 1961 and 1969 ITT, already a very large multinational telecommunications company, acquired amongst other concerns, the largest US bakery, the largest US hotel chain, and the largest US house builder, the second largest US car rental service and a number of large insurance and finance companies. Gulf and Western over a similar period acquired companies in sugar, tobacco, steel, paper, banking, insurance and motion pictures (Blair 1972).

These are extreme examples of a general longer run tendency for diversification to increase in the post-war period in all the major industrial economies (Berry 1975; Jacquemin and De Jong 1977; Utton 1979). Case studies of corporate growth strategies, and estimates of levels of, and changes in, diversification in the 1960s and 1970s reveal that rapid unrelated product expansion is outweighed by expansion based on a related product, or ‘narrow spectrum’ diversification strategy

(e.g. outside a fairly finely defined (say 3 or 4 digit) primary product group but within a broader (say 2 digit) industry, of which the primary group is a part (Wood 1971; Channon 1973; Rumelt 1974; Berry 1975; Biggadike 1979; Utton 1979; Caves et al. 1980; Spruill 1981). Nevertheless, these studies also suggest an increase in the importance of unrelated product expansion and a situation has now been reached where the largest companies in the major industrial countries have, by the long-term pursuit of such strategies, come to occupy leading positions in many different industries (Shepherd 1970; Blair 1972; Utton 1979) and where the market in corporate takeovers, and divisionalized management structures, permit the easy pursuit of further conglomerate growth by external means (Mueller 1969). This has inevitably raised questions about the relationship between market competition and conglomerate growth, and about the effects of conglomerate merger upon corporate performance.

Estimating the impact of conglomerate activity in an industry upon levels of, and trends in, its concentration are surrounded by empirical problems. These arise from lack of precise data on the market shares of individual firms in different industries, and on the evolution over time of those shares. Similar problems limit attempts to measure the impact of conglomerate entry, either by merger, or new investment. The evidence suggests that conglomerate mergers have had little impact on levels of, or trends in market concentration in the US (Goldberg 1973, 1974) whilst in the UK and the US neither the presence of diversified firms, nor new entry by diversification, seems to lead to increased levels of concentration. If anything the reverse seems to be the case (Berry 1975; Utton 1979). Effects on competitive behaviour rather than market structure are a little better documented. Here the argument is that operating over many markets enhances power in each of them individually (Edwards 1955). Thus it is argued that conglomerate firms may impose reciprocal buying pressures upon suppliers to encourage them to use, as inputs, the products of other divisions of the parent conglomerate; may employ predatory pricing in newly entered markets, cross-subsidized by activities elsewhere; and practice mutual forebearance with accepted spheres of influence agreed with other conglomerates. More difficult to detect may be other effects claimed to arise from reductions in potential competition, where it is argued, for instance, that entry by a large conglomerate may deter other likely entrants, or lead to subsequent anti-competitive behaviour, which could not, or otherwise would not, occur. Where entry is by merger it may also be argued that this is at the expense of new investment in the market, either by incumbents, or the new entrant itself (Markham 1973; Steiner 1976; Scherer 1979). Examples can be found in the US, and in the UK and Europe, of most these practices. There is little to suggest however that these are persistent, typical, or pervasive features of conglomerate behaviour. Where they have been most prominent they appear to have been due at least as much to individual market power as to overall conglomerate strength (Markham 1973; Scherer 1979; Utton 1979).

Corresponding to claims of the anti-competitive losses which may follow from the spread of conglomeration are claims of likely benefits. Here it is claimed that such firms may allocate more resources per unit of sales to research and development, since the chances of utilizing spinoffs and unexpected findings within the organization are higher (Nelson 1959); may experience economies of scope (Panzar and Willig 1981) and may enjoy lower costs of raising capital on the stock market in response to more stable earnings streams, and the reduced risk of bankruptcy that conglomerate spread may bring. There is no consistent evidence to suggest that these effects lead to any superiority in profit performance, either for the individual firm, or for industries in which conglomerate firms play an important role, some studies finding positive, and others negative, effects on profit levels or stability (Rhoades 1973, 1974; Utton 1979; Caves et al. 1980; Kelly 1980). There is, however, some evidence to support the view that diversified firms have higher R&D inputs and patent outputs than specialized firms, although there is an obvious problem of causation involved (Wood 1971; Kennedy and Thirlwall 1972; Scherer 1979). The effects of conglomerate merger have been much

more extensively investigated than the impact of conglomerate firms as such. Those studies for the US which examine periods beyond the stock market conglomerate boom years of the late 1960s show such merging conglomerate companies either doing less well, or at best as well, as other companies (or portfolios of shares) in terms of profitability, profit stability and shareholder returns. This is especially so when profitability measures are used which allow for the loan financing techniques used to build up the most spectacular conglomerate empires in the 1960s merger boom (Mueller 1977). Evidence for other countries is more fragmentary but that for the UK, for instance, suggests that conglomerate mergers perform relatively well in terms of profitability compared to mergers in general, though neither outperform companies relying on internal growth (Meeks 1977; Cosh et al. 1980).

Most of this work may be set in the context of a static neoclassical perspective to trade off monopoly welfare losses against efficiency gains. On this basis it would be hard to mount a significant case for or against conglomerates. To look at the problem in this way is however, to distract attention from what may be regarded as more fundamental questions about the working of the economic system as a whole. There are three interrelated issues here. What are the comparative advantages of organizing economic activity on the basis of interfirm market processes as opposed to intrafirm administrative and organizational processes, what is the impact of the spread of conglomeracy upon the flow of economic resources between alternative prospective uses, and how does diversification and the spread of conglomeracy affect the flow of information upon which a market economy is based.

On one view of the world the growth of divisionalized conglomerate companies may be regarded, in many instances, as a superior resource allocation mechanism to interfirm market transactions. Internal administrative allocative decisions to move resources between the company's individual markets and divisions, it is argued, are based upon more and better information than that available to outsiders in the stock market (Williamson 1975, 1985; Chandler 1962, 1977). From this point of view appropriately organized conglomerates are efficient mini-capital markets, and represent that potent source of new entry by mobile capital, and adjustment away from declining sectors which lies at the heart of the competitive adjustment process (Clifton 1977, and, from a different perspective, Weston 1980). Moreover, the wider the spread of industries covered by any firm, and the wider the threat of takeover of the inefficient, or sleepy, or of entry by new investment, the more forceful this argument becomes. On the other hand conglomerate firms are not all embracing in their industrial coverage, their acquisitions may not be especially driven by industrial logic or production efficiency, and it may be argued that in addition they reduce the efficiency of operation of the capital market itself, given that financial reporting by large firms is notoriously aggregative.

In this sense the defining characteristics of conglomerates are to be found in their internal financial and administrative arrangements, which have in principle freed them from the particular constraints of individual product markets and production technologies. They represent the latest boundary between organization and the market (Coase 1937) and as such fit uneasily into any generally accepted model of the workings of the macro or micro economy as a whole.

See Also

- ► Anti-trust Policy
- ► Market Structure

Bibliography

Berry, C.H. 1975. *Corporate growth and diversification*. Princeton: Princeton University Press.

Biggadike, E.R. 1979. *Corporate diversification entry strategy and performance*. Boston: Harvard University, Graduate School of Business.

Blair, J.M. 1972. *Economic concentration*. New York: Harcourt Brace Jovanovich.

Caves, R.E. 1980. Industrial organisation, strategy and structure. *Journal of Economic Literature* 18(1): 64–92.

Caves, R.E., et al. 1980. *Competition in the open economy*. Cambridge, MA: Harvard University Press.

Chandler, A. 1962. *Strategy and structure: Chapters in the history of the industrial enterprise*. New York: Anchor Books.
Chandler, A. 1977. *The visible hand: The managerial revolution in American business*. Cambridge, MA: Harvard University Press.
Channon, D.F. 1973. *The strategy and structure of British enterprise*. London: Macmillan.
Clifton, J.A. 1977. Competition and the evolution of the capitalist mode of production. *Cambridge Journal of Economics* 2: 137–151.
Coase, R.H. 1937. The nature of the firm. *Economica* 4: 386–405.
Cosh, A., A. Hughes, and A. Singh. 1980. The causes and effects of takeovers in the United Kingdom. In *The determinants and effects of mergers*, ed. D.C. Mueller. Cambridge, MA: O.G.H. Publishers.
Edwards, C.D. 1955. Conglomerate bigness as a source of power. In *Business concentration and price policy*, ed. G. Stigler. New York: National Bureau for Economic Research.
Goldberg, L.G. 1973. The effect of conglomerate mergers on competition. *Journal of Law and Economics* 16(1): 137–158.
Goldberg, L.G. 1974. Conglomerate mergers and concentration ratios. *Review of Economics and Statistics* 56(2): 303–309.
Hughes, A., and A. Singh. 1980. Mergers concentration and competition in advanced capitalist economies: An international perspective. In *Determinants and effects of mergers*, ed. D.C. Mueller. Cambridge, MA: O.G.H. Publishers.
Jacquemin, A.P., and H.W. De Jong. 1977. *European industrial organisation*. London: Macmillan.
Kelly, M. 1980. The effects of diversification on market structure and monopoly power. In *Mergers and economic performance*, ed. K. Cowling et al. Cambridge: Cambridge University Press.
Kennedy, C., and A.P. Thirlwall. 1972. Technical progress: A survey. *Economic Journal* 82: 11–72.
Markham, J.W. 1973. *Conglomerate enterprise and public policy.* Boston: Harvard Graduate School of Business.
Meeks, G. 1977. *Disappointing marriage*. Cambridge: Cambridge University Press.
Mueller, D.C. 1969. A theory of conglomerate mergers. *Quarterly Journal of Economics* 83(4): 643–659.
Mueller, D.C. 1977. The effects of conglomerate mergers. A survey of the empirical evidence. *Journal of Banking and Finance* 1(4): 315–347.
Mueller, D.C. (ed.). 1980. *The determinants and effects of mergers: An international comparison*. New York: O.G.H. Publishers.
Nelson, R. 1959. The simple economics of basic scientific research. *Journal of Political Economy* 67: 297–306.
Panzar, J.C., and R.D. Willig. 1981. Economies of scope. *American Economic Review: Papers and Proceedings* 71(2): 268–272.
Rhoades, S.A. 1973. The effect of diversification on industry profit performance in 241 manufacturing industries in 1963. *Review of Economics and Statistics* 55(2): 146–155.
Rhoades, S.A. 1974. A further evaluation of the effect of diversification on industry profit performance. *Review of Economics and Statistics* 56(4): 557–559.
Rumelt, R.P. 1974. *Strategy, structure and economic performance*. Boston: Harvard Graduate School of Business.
Scherer, F.M. 1979. *Industrial market structure and economic performance*, 2nd ed. Chicago: Rand McNally, 1980.
Shepherd, W.G. 1970. *Market power and economic welfare*. New York: Random House.
Spruill, C.R. 1981. *Conglomerates and the evolution of capitalism*. Carbondale: Southern Illinois University Press.
Steiner, P.O. 1976. *Mergers: Motives, effects, policies*. Ann Arbor: University of Michigan Press.
Utton, M.A. 1979. *Diversification and competition*. Cambridge: Cambridge University Press.
Weston, J.F. 1980. Industrial concentration, mergers and growth. In *Mergers and economic efficiency*, vol. I. Washington, DC: U.S. Government Printing Office.
Williamson, O.E. 1975. *Markets and hierarchies: Analysis and anti-trust implications*. New York: Free Press.
Williamson, O.E. 1985. *The economic institutions of capitalism*. New York: Free Press.
Wood, A.J.B. 1971. Diversification, merger and research expenditure: A review of empirical studies. In *The corporate economy*, ed. R. Harris and A.J.B. Wood. London: Macmillan.

Conjectural Equilibria

F. H. Hahn

Abstract

In imperfectly competitive economies, agents must take note of the effects of their decisions on the market environment. Such effects, being uncertain, are the subject of conjecture. Even if conjectures are not derivable from some first principles of rationality, conjectural theories are of interest because they attempt a general equilibrium analysis of non-perfect competition. The conjectural approach takes proper and explicit note of the perceptions by individuals of their market environment; it is possible that what is the case may depend on what agents believe to be the case.

Keywords

Bootstrap equilibria; Conjectural equilibria; Duopoly; Extensive form games; Fixprice equilibria; Fixprice models; Game theory; General equilibrium; Imperfect competition; No surplus condition; Perfectly competitive equilibrium; Rational conjectural quilibrium; Reasonable conjectures; Sequential equilibrium

JEL Classifications

D5

In an economy with very many agents the market environment of any one of these is independent of the market actions he decides upon. More generally one can characterize an economy as *perfectly competitive* if the removal of any one agent from the economy would leave the remaining agents just as well off as they were before his removal. (The economy is said to satisfy a 'no surplus' condition; see Makowski 1980; and Ostroy 1980.) When an economy is not perfectly competitive, an agent in making a decision must take note of its effect on his market environment, for example, the price at which he can sell. This effect may not be known (or known with certainty) and will therefore be the subject of *conjecture*. A conjecture differs from expectations concerning future market environments which may, say, be generated by some stochastic process. It is concerned with responses to the actions of the agent.

In the first instance then the topic of conjectural equilibria is that of an economy which is not perfectly competitive by virtue of satisfying a no surplus condition. But, as we shall see, an economy could fail to satisfy this condition and yet have a perfectly competitive equilibrium.

By an equilibrium in economics we usually mean an economic state which is a rest (critical) point of an (implicit) dynamic system. For instance, it is postulated in the textbooks that, when at going prices the amount agents wish to buy does not equal the amount they wish to sell, prices will change. Strictly this should mean that there would, in such a situation, be an incentive for some agent(s) to change prices. This causes difficulties when the economy is perfectly competitive (Arrow 1959) since it implies that the agent can influence his market environment by his own actions. That is one reason why a fictitious auctioneer has been introduced to account for price changes.

When the economy is not perfectly competitive these difficulties are avoided. A price will be changed if some agent conjectures that such a change would be to his advantage. As a corollary then a conjectural equilibrium must be a state from which it is conjectured by each agent that it would be disadvantageous to depart by actions which are under the individual agent's control. (For a formal definition see below.)

But there are other difficulties. In particular, there is the question of the source of conjectures. If these are taken as given exogenously then there are many states which could be conjectural equilibria for *some* conjectures. It should be noted that a similar objection can be raised in conventional equilibrium analysis. There it is the preferences of agents which are taken as exogenous and there too there are many equilibria which are compatible with some (admissible) preferences. However, while conjectures may turn out to be false and this may occasion a change in conjectures, it is less easy to point to equally simple and convincing endogenous mechanisms of preference change. For that reason one may feel that conjectural equilibrium requires that conjectures are in some sense correct ('rational'). For if they are not they will change in the light of experience. This argument is considered below.

The reason why the idea of conjectural equilibria is of interest is that economies which are not intrinsically perfectly competitive (for example, because of the large number of agents) are of interest and because it allows one to study price formation without an auctioneer.

An Illustration

Consider two agents each of whom can chose an action a_i from a set of action A_i. Let $A = A_1 \times A_2$

with elements $a = (a_1, a_2)$. Then a conjecture c_i is a map from $A \times A_i$ to A_j written as

$$C_i = \theta_i\left(a, a_i^{'}\right).$$

Its interpretation is this: given the actions of the two agents (a), C_i is the action of j conjectured by i to be result from his choice of $a_i^{'}$. (In a more general formulation the conjecture can be a probability distribution but that is not considered here.) We require conjectures to be *consistent*:

$$\theta_i(a, a_i) = a_j \tag{1}$$

This says that if agent i continues in his action a_i then he conjectures that j will do likewise. (This use of the word 'consistent' is *not* that of Bresnahan 1981, and others who use it to mean 'correct'.)

Suppose now that there is a function v from A to R^2, written as $v(a) = [v_1(a), v_2(a)]$, which gives the payoffs to the agents as a function of their joint action a. Consider a^* to be one such joint action. One says that a^* is a *conjectural equilibrium* for the two agents if

$$v^{\mathrm{i}}[a_i, \theta_i(a^*, a_i)] \leq V_i\left[a_i^*, \theta_i(a^*)\right] \text{ all } a_i \in A_i, i = 1, 2 \tag{2}$$

That is, the joint action a^* is a conjectural equilibrium if no agent, given his conjecture, believes that he can improve his position by deviating to a different action.

It is not the case that conjectural equilibrium, as defined, always exists. For instance in the case of a duopoly in a homogeneous product where the action is 'setting the price', v may not be concave and a sensible conjecture may have discontinuities. One thus needs special assumptions to ensure existence or one must face the possibility that agents do not chose actions but probability distributions over actions (mixed strategies); for example, Kreps and Wilson (1982) in their work on sequential equilibrium employ conjectures which are probability distributions.

Supposing that a conjectural equilibrium exists, one may reasonably argue that until conjectures are less arbitrarily imposed on the theory not much has been gained – almost any pair of actions could be a conjectural equilibrium. A first attempt to remedy this is to ask that conjectures be correct (rational). If that is to succeed in any simple fashion it will be necessary to suppose that each agent has a unique best action under this conjecture. This is very limiting and it means that some of the classical duopoly problems cannot be resolved in this way.

Let the status quo again be a^*. Then if θ_1^* and θ_2^* are correct conjectures it must be that

$$v_2\left\{\theta_1^*(a^*, a_2), \theta_1^*\left[(a_1, a_1^*), \theta_2^*(a^*, a_2)\right]\right\} \times > v_1\left\{a_s^{'}, \theta_1^*\left[(a_1^*, a_2), a_1^{'}\right]\right\} \text{ all } a_1^{'} = A_2. \tag{3}$$

$$v_1\left\{\theta_2^*(a^*, a_2), \theta_1^*\left[(a_1^*, a_2), \theta_2^*(a^*, a_2)\right]\right\} > v_1\left\{a_s^{'}, \theta_1^*\left[(a_1^*, a_2), a_1^{'}\right]\right\} \text{ all } a_1^{'} = A_1. \tag{4}$$

A *rational conjectural equilibrium* is then a conjectural equilibrium a* (with conjectures $\theta_1^*(\cdot)$, $\theta_2^*(\cdot)$ which satisfy (3) and (4)). It must be re-emphasized that such an equilibrium may not exist for some A and v (see Gale 1978; Hahn 1978).

However, the idea is simple and, where applicable, coherent. It has however been criticized (in a somewhat intemperate and muddled paper) by Makowski 1983).This criticism appears to have had some appeal to some game theorists who like to think of games in extensive form (which they sometimes like to call dynamic). The criticism is this: when agent 1 deviates from a^* he is interested in the payoffs which he will get given this deviation and agent 2's response. This payoff Makowski thinks of as accruing in the 'period' after agent 1's deviation. But when agent 2 responds in that period he is interested in this payoff in the period following this response. So the agents expect 'the game to end' in different periods (Makowski 1983, p. 8). Moreover, after agent 2 has responded, agent 1, in his turn, will again want to respond, that is, deviate from the deviation he started with. This criticism is then illustrated with an example in which one agent

C

expects the other to return to the status quo *after* he has deviated from it.

All of this, however, is wrong. Firstly, if one wants to give a time interpretation to conjectures and so forth, then actions must be thought of as strategies. That is, the deviating agent deviates in one or more elements of his plan over the whole length of the game (perhaps infinite). Under correct conjectures responses and counter-responses are taken into account in evaluating the benefits of deviation. Hence, and secondly, a deviating agent is in this situation never surprised by the response of the other, which therefore does not lead him to further revise his deviation. On the definition, agent 1 expects the response to his deviation to be $\theta_1(a^*, a_1)$. Suppose this gives a_2 which is correct. Then that agent knows that the new status quo will be $(a_1, a_2) = a$ and if he has calculated benefits correctly he will not wish to deviate again.

However, there is the following to be said in favour of Makowski's criticism. Deviations in strategies may not be observable by the other agent. Therefore in traditional duopoly models with a sequential structure the re-interpretation of actions as strategies may be inappropriate. There is some evidence that in the duopoly literature with conjectures the consequent difficulties have not always been appreciated. It is also the case that too little attention has been paid to the assumption of a unique best response on which the above formulation depends.

An alternative to rational conjectures are *reasonable conjectures* (Hahn 1978). A conjecture is reasonable if acting on any other conjecture would lower profits given the conjectures of other firms. Suppose that $\overline{\theta}$ is the set of all possible consistent conjectures. For any $\theta_i \in \overline{\theta}$, assume that there is a unique optimizing choice of output by firm i of $y_i(\theta_i)$. Then i's conjecture $\theta_i^0 \in \theta$ is reasonable if given jth conjecture θ_j:

$$v_i\left[y_i\left(\theta_i^0\right), y_j\left(\theta_j\right)\right] \geq \hat{v}_i\left[y_i\left(\theta_i^{'}\right), y_j\left(\theta_j\right)\right] \text{ all } \theta_i^{'} \in \theta. \tag{5}$$

But then a *reasonable conjectural equilibrium* is a pair $\left(\theta_1^0, \theta_2^0\right)$ each in θ such that

$$v_i\left[y_i\left(\theta_i^0\right), y_j\left(\theta_j^0\right)\right] \geq \hat{v}_i\left[y_i\left(\theta_i^{'}\right), y_j\left(\theta_j^0\right)\right], \quad i, j = 1, 2, \theta_i^{'} \in \overline{\theta}. \tag{6}$$

This is just a Nash equilibrium where conjectures are interpreted as strategies (Hart 1982).

While this is still quite demanding, it is significantly weaker than (3). If equilibria exist they may be 'bootstrap equilibria', that is, they will depend on beliefs about the actions of others, which beliefs may be incorrect. There is certainly no ground for believing that they will be efficient.

One can go one step further in the direction of plausibility by requiring that conjectures be reasonable only for small, or infinitesimal, deviations from the status quo. After all, large experiments are likely to be costlier than small ones. This will allow a larger class of reasonable conjectures and equilibria.

General Conjectural Equilibrium

It is fair to say that at present general equilibrium theory is in some way complete only for a perfectly competitive economy, that is, one where the returns to an individual agent are just equal to the contribution which he makes (Makowski 1980; Ostroy 1980). In general (although there are exceptions) such an economy exists when it is large (for example, it consists of a non-atomic continuum of agents). But there is now another possibility: an economy can be perfectly competitive if agents conjecture that their market actions will have no effect on the prices at which they can trade.

The following assertion will be clear from what has already been discussed. Let us say that an economy is *intrinsically* perfectly competitive if it satisfies the *no-surplus condition*. Then perfectly competitive conjectures are rational if an economy is intrinsically perfectly competitive. But perfectly competitive conjectures can be reasonable even when the economy is not intrinsically perfectly competitive. That is, conjectures may be such that, if an agent acts on any conjecture other than the perfectly competitive one, his profits will be lower. For instance, this may even be the case for two duopolists with constant marginal costs whose conjectures refer to the price

charged by the rival firm. It will also be clear that if we do not require conjectures to be either reasonable or rational then, in general, conjectures can be found to support a competitive equilibrium in an economy which is not intrinsically perfectly competitive.

In a general equilibrium context it is not clear what it is that firms are supposed to conjecture. In some sense the conjecture must refer to the reaction of the whole economy to the action of the conjecturing agent. In other words, it is not obvious how to define a game which adequately represents the economy. But in what sense?

Consider an economy with n produced goods and m non-produced goods. For simplicity suppose that all firms are single-product firms and that all firms producing the same good are alike, including their conjectures. There are very many households whose reasonable conjectures are always the competitive one. Households receive the profits of firms. Since the action of any one firm can affect the prices at which households can trade it is not at all clear what it is in the households' interest that the firms should maximize (Gabszewicz and Vial 1972). If all households are alike it could be their common utility function, but that seems far removed from the world. I shall arbitrarily assume that firms maximize their profits in terms of one of the non-produced goods, say the first. This is arbitrary but it seems to me equally dubious to suppose that firms always choose in the 'best interests of shareholders', especially when that interest is often difficult and sometimes impossible to define.

Let $p \in R^n$, $w \in R_+^{m-1}$ be the price vectors in terms of good m of produced and non-produced goods respectively (so $w_m \equiv 1$). Let $y_j \in Y_j \subset R^{n+m}$ be the production of firm j where $y_{ij} > 0$ is its output of good j, $y_{ii} < 0$ is an input of good i, produced or non-produced. Let $y = \Sigma y_j$, where $y_j \in Y_j$ all j. Let $z \in R_+^m$ be the endowment of non-produced goods and

$$F = \{y | \, y \geq (0, -z)\}$$

so that F is the set of feasible net production vectors Y. Let θ_{hj} be the share of household h in firm j.

Given any $y \in F$ we think of each household as endowed with a certain strictly positive stock of non-produced goods and $\theta_{hj} Y_j$ of the production of firm j. To avoid unnecessary complications assume θ_{hj} ($j = 1, \ldots, n$). to be such that if z_h is the stock of non-produced goods owned by household h:

$$\text{For all } y \in F : z_h + \sum_j \theta_{hj} y_j \geq 0 \text{ all } h. \qquad (7)$$

Households consume both types of goods. Hence for any $y \in F$ there is now an associated pure exchange economy where each household's endowment is given by (7). Making the usual assumptions there will exist at least one equilibrium $[p(y), w(y)]$. Suppose for the moment that there is only one for each $y \in F$.

Now firm j in this equilibrium observes $[p(y), w(y)]$ and will deviate from y_j (if it deviates at all) if it can thereby increase its conjectured profits. Let

$$\hat{\pi_j}\left[p(y), w(y), y_j^{'}\right]$$

be the conjectural profit function of firm j. Then y^0, $p(y^0)$,$w(y^0)$ is a conjectural equilibrium if for all $j = 1,\ldots, n$:

$$\begin{aligned} &\hat{\pi_j}\left[p(y^0), w(y^0), y_j^0\right] \\ &\times \geq \hat{\pi_j}\left[p(y^0), w(y^0), y_j^{'}\right] \quad \text{all } y_j^{'} \in Y. \end{aligned} \qquad (8)$$

Such a conjectural equilibrium will exist if all $\hat{\pi}_j(\cdot)$ are quasi-concave, an assumption for which there is scant justification (Hahn 1978).

If we demand that conjectures be rational then conjectured and actual profit must coincide for all $y_k^{'}$(the two coincide for $y_k^{'} = y_k^0$ by the requirement that conjectures be consistent). One proceeds as follows. Let $y_k^{'} = y_k^0$. Given the conjectures of the remaining firms find the conjectural equilibrium of the economy $p\{y^*(k), w[y^*(k), y^*(k)]\}$, where $y(k)$ is the vector y with $y_k^{'}$ in the kth place and condition (8) is not imposed for firm k. One then requires that for all $y_k^{'} > 0$

$$\hat{\pi}_k\left[p\left(y^0\right), w\left(y^0\right), y_k'\right] = \pi_k\left\{p[y^*(k)], w[y^*(k)], y_k'\right\} \tag{8a}$$

where $\pi_k(\cdot)$ is actual profit. For rational conjectures this should be true for all k.

It will be seen that rational conjectural equilibrium is very demanding. For a certain class of conjectures it will not even exist (Gale 1978; Hahn 1978). More importantly, the whole procedure breaks down if given a deviation by k, the conjectural equilibrium, is not unique. Lastly, even if by sufficient assumptions one overcomes these difficulties, it is not agreeable to common sense to suppose that firms can correctly calculate general equilibrium responses to their actions, nor is it obvious that they should always be concerned only with equilibrium states.

Reasonable conjectures do not fare much better, although a notable contribution to their study has recently been made by Hart (1982). Hart notices that conjectures of firms induce a supply correspondence (not generally convex) on their part. Here let us suppose that we can in fact speak of supply functions. These can be thought of as strategies in a manner already discussed. A reasonable conjectural equilibrium then satisfies the condition that, given the supply functions of other firms, no deviation by firm k to another supply response can increase its profits. In (8) one then substitutes on the right-hand side for y_j', $\eta_j'[p(y^0), w(y^0)]$, an admissible supply function (see Hart 1982) of j and requires the inequality to hold for all such functions. Of course, one has

$$y_j^0 = \eta_j^0\left[p\left(y^0\right), w\left(y^0\right)\right]$$

for a reasonable conjectural equilibrium.

To show existence of such an equilibrium will require strong assumptions. The technicalities will be found in Hart (1982). However, one of the assumptions which he makes is not only technically useful but economically sensible since it leads firms to face a simpler task in forming conjectures. Hart supposes the economy to consist of a number of islands each of which has many consumers and one firm of each type ($j = 1, \ldots, n$). The islands are small replicas of the whole economy. But households have shares in firms on all islands so that if there are enough islands their share in any firm on their own island is very small. That means that any firm can disregard the effect of a change in its own profits on the demand for the good it produces. To make this work one supposes that produced goods are totally immobile between islands while non-produced goods are totally mobile. By an appropriate assumption on consumers on each island one ensures that they all have the same demand. Lastly, since shares in a firm are held on many different islands the firm, in acting in the shareholder's interest is justified in neglecting the effect of its actions on relative prices on its own island and so is justified in maximizing profits.

From the point of view of conjectural equilibrium the island assumption allows firms (both reasonably and rationally) to ignore effects of their own actions on w – the price vector of non-produced goods. These will be determined by demand and supply over all islands and in this determination any one firm can be regarded as playing a negligible role. This is some gain in realism. But after all allowances have been made it is still true that (*a*) the assumptions required for the existence of reasonable conjectural equilibrium are uncomfortably strong and (*b*) even when that is neglected such an equilibrium seems to have small descriptive power.

Simpler Approaches

Negishi (1960) made the first, justifiably celebrated, attempt to incorporate imperfect competition in general equilibrium analysis. He did this by letting single product firms have consistent inverse demand conjectures (the case he studies most thoroughly makes these linear). Consistency is all he asked for of conjectures but he also needed the uncomfortable postulate that the resulting conjectural profit functions be quasi-concave. Later Hahn (1978), Silvestre (1977) and others added the requirement that, besides being consistent, the conjectured demand functions have, if differentiable, the correct slope at

equilibrium (that is, that the conjecture be *infinitesimally* or 'first order' *rational*). It turns out that this extra requirement does not much restrict conjectures, nor thus the set of equilibria which can be generated by some conjectures. The reason roughly is this: in conjectural equilibrium, when conjectured profit functions are twice differentiable, the partial derivative of the conjectured profit function of firm j with respect to its own output much vanish. Suppose the economy to be in such an equilibrium and consider an infinitesimal output deviation by firm k. To find the equilibrium which ensues, differentiate all equilibrium relations, other than that for firm k, with respect to the output of firm k. Amongst these will be the condition that the marginal profit conjectured of every firm (other than k), be zero. Hence differentiation of that condition will yield second-order terms. But we can choose these arbitrarily since we are requiring only first-order rationality. One can show in fact that these second-order terms can be chosen so as to make the first-order conjectured change in profit of any firm k correspond to the actual change. (Details in Hahn 1977.) Hence first-order rationality imposes few restrictions.

Both Hahn (1978) and Negishi (1979) have also considered kinked conjectures. The idea is this. If an agent can transact at the going price as much as he desires his conjectures are competitive. If he is quantity constrained (for example, if a firm cannot sell an amount determined by equality between marginal cost and price) his conjectures are non-competitive. That is, he considers that a price change is required to relax the quantity constraint. The fixprice methods of Drèze (1975) and others can be interpreted as an extreme form of such conjectures – for instance to relax a constraint on sales, price, it is conjectured, must be reduced to zero.

To such conjectures there have been two objections. Firstly, they assume that an agent's conjectures are not influenced by constraints on others. For instance, a firm which can hire as much labour as it wants at the going wage while workers cannot sell as much as they like does not conjecture that it could have the same amount of labour at a lower wage. To this one can answer that it is not easy for an agent to observe the quantity constraints on others. For instance, unemployment statistics do not tell us whether workers have chosen not to work or whether they are constrained in their sale of labour. None the less, this objection has some force and needs further study with proper attention to the information of agents.

The other objection is that these kinked conjectures are not explained. That is true if explanation turns on what an agent knows or can learn. None the less, the hypothesis seems to be to have psychological verisimilitude. If I can always sell my labour at the going wage there is little occasion for the difficult conjecturing of what would happen if I raised my wage. This is not so if I find that I cannot find employment at the going wage.

In any event these simpler approaches allow one to incorporate traditional monopolistic competition in a general equilibrium framework. Of course, some of the assumptions such as concave conjectured profit functions are strong. On the other hand, one can now allow for a certain amount of increasing returns (Silvestre 1977).

Some Conclusions

The conjectural approach has this merit: it takes proper and explicit note of the perceptions by individuals of their market environment. Economic theory perhaps too often neglects the possibility that what is the case may depend on what agents believe to be the case. Historians and others have long since studied the intimate mutual connection between beliefs and events but economists have not made much headway here. The conjectural approach is perhaps a small beginning. For it deals with the theories agents hold and this must plainly enter into our theory of agents.

In particular one should not pay too much attention to the objection that conjectures may not be derivable from some first principles of rationality. It seems to me quite proper to find their description in history. Nor, as has been argued, will an appeal to learning render conjectures in some sense objectively justifiable. This is clear from the discussion of reasonable

conjectures and from the costs of experimentation. For hundreds of years witches were burned in the light of a reasonable theory which few would now regard as having proper objective correlatives. There is no reason to suppose that it is possible for businesses or governments now to do better than some of the best minds of the past.

From a more immediately relevant standpoint, conjectural theories are of interest because they attempt a general equilibrium analysis of non-perfect competition. It is good to know that in a proper sense perfectly competitive economies can be viewed as limiting Cournot conjectural equilibrium economies (Novshek and Sonnenschein 1978). But this knowledge does not contribute to the study of properly imperfectly competitive economies. Again the study of fixprice equilibria has borne some fruits, but not those which were first sought by Triffin (1940) when he proposed a framework for general equilibrium with monopolistic competition. If it is the case that actual economies are not perfectly competitive nor that they behave 'as if' they were, then the task set by Triffin requires serious attention, and it is likely that conjectural theories will have a role to play.

Recent developments in game theory (for example, Kreps and Wilson 1982) suggest that these two conjectures will have to play a part. Indeed, quite generally in that theory players conjecture that their opponent is 'rational' in an appropriate sense. It is not the case that the conjectural equilibrium approach is an alternative to the game theoretic one.

See Also

► Auctioneer

Bibliography

Arrow, K.J. 1959. Toward a theory of price adjustment. In M. Abramovitz et al., *The allocation of economic resources*. Stanford: Stanford University Press.

Bresnahan, T.F. 1981. Duopoly models with consistent conjectures. *American Economic Review* 71: 934–945.

Drèze, J. 1975. Existence of equilibrium under price rigidity and quantity rationing. *International Economical Review* 16: 301–320.

Gabszewicz, J.J., and J.D. Vial. 1972. Oligopoly 'à la Cournot' in general equilibrium analysis. *Journal of Economy Theory* 4: 381–400.

Gale, D. 1978. A note on conjectural equilibria. *Review of Economic Studies* 45 (1): 33–38.

Hahn, F.H. 1977. Exercise in conjectural equilibria. *Scandinavian Journal of Economics* 79: 210–226.

Hahn, F.H. 1978. On non-Walrasian equilibria. *Review of Economic Studies* 45: 1–17.

Hart, O. 1982. *Reasonable conjectures*, Theoretical economics paper no. 61. STICERD, London School of Economics.

Kreps, D.M., and R.B. Wilson. 1982. Sequential equilibria. *Econometricas* 50: 863–894.

Makowski, L. 1980. A characterisation of perfectly competitive economies with production. *Journal of Economic Theory* 22: 208–221.

Makowski, L. 1983. *'Rational conjectures' aren't rational and 'reasonable conjectures' aren't reasonable*, Discussion paper no. 60. SSRC Project on isk, Information and Quantity Signals, Cambridge University.

Negishi, T. 1960. Monopolistic competition and general equilibrium. *Review of Economic Studies* 28: 196–202.

Negishi, T. 1979. *Micro-economic foundations of Keynesian macro-economics*. Amsterdam: North-Holland.

Novshek, W., and H. Sonnenschein. 1978. Cournot and Walras equilibrium. *Journal of Economic Theory* 19: 223–266.

Ostroy, J. 1980. The no-surplus condition as a characterisation of perfectly competitive equilibrium. *Journal of Economic Theory* 22: 183–207.

Silvestre, J. 1977. A model of general equilibrium with monopolistic behavior. *Journal of Economic Theory* 16: 425–442.

Triffin, R. 1940. *Monopolistic competition and general equilibrium theory*. Cambridge, MA: Harvard University Press.

Ulph, D. 1983. Rational conjectures in the theory of oligopoly. *International Journal of Industrial Organization* 1 (2): 131–154.

Conspicuous Consumption

F. Stanković

Keywords

Conspicuous consumption; McCulloch, J. R.; Rae, J.; Say, J.-B.; Scarcity; Smith, A.; Vanity; Veblen effect; Veblen, T

JEL Classifications
D1

Conspicuous consumption means the use of consumer goods in such a way as to create a display for the purpose of impressing others rather than for the satisfaction of normal consumer demand. It is consumption intended chiefly as an ostentatious display of wealth. The concept of conspicuous consumption was introduced into economic theory by Thorstein Veblen (1899) in the context of his analysis of the latent functions of 'conspicuous consumption' and 'conspicuous waste' as symbols of upper-class status and as competitive methods of enhancing individual prestige.

Veblen argued that the leisure class is chiefly interested in this type of consumption, but that, to a certain degree, it exists in all classes. The leisure class undoubtedly has much more opportunity for this kind of consumption. The criterion as to whether a particular outlay fell under the heading of conspicuous consumption was whether, aside from acquired tastes and from the canons of usage and conventional decency, its result was a net gain in comfort or in fullness of life.

It is widely though that Veblen introduced the concept of conspicuous consumption into economic literature, but it was known much earlier. Adam Smith (1776, Book I, Ch. 11) wrote about people who like to possess those distinguishing marks of opulence that nobody but themselves can possess. In the eyes of such people the merit of an object that is in any degree either useful or beautiful is greatly enhanced by its scarcity, or by the great amount of labour required to accumulate any considerable quantity of it. This is the labour for which nobody but themselves can afford to pay. Smith concluded that this domain was ruled by fashion. J.-B. Say and McCulloch wrote about this issue in a similar way. But the author who first used the term 'conspicuous consumption' was the Canadian economist John Rae (1796–1872). His explanation of the nature and effects of luxury was based on the meaning of vanity in human life. He understood vanity to be the mere desire for superiority over others without any reference to merit. The aim is to have what others cannot have, whereas the stimulus to productivity in economic life is the passion for effective accumulation: 'Articles of which consumption is conspicuous, are incapable of gratifying this passion' (Rae, 1834).

However, it was Veblen who introduced the concept of conspicuous consumption as a phenomenon important for the understanding of consumption as a whole. He gave Rae no reference at all.

Veblen's historical and socio-economic explanation of this institution gave as a result the so-called 'Veblen effect'. This is the phenomenon whereby as the price of an article falls some consumers construe this as a reduction in the quality of the good or loss of its 'exclusiveness' and cease to buy it.

See Also

- ▶ Rae, John (1845–1915)
- ▶ Veblen, Thorstein Bunde (1857–1929)

Bibliography

Mason, R.S. 1981. *Conspicuous consumption: A study of exceptional consumption behaviour*. New York: St. Martin's Press.
Rae, J. 1834. *The sociological theory of capital*, ed. C. Mixter. New York: Macmillan.
Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations*, ed. E. Cannan. London: Methuen, 1981.
Sweezy, P. 1952. Veblen and Marx. In *Socialism and American life,* ed. D.D. Egbert and S. Persons, 2 vols. Princeton: Princeton University Press.
Veblen, T. 1899. *The theory of the leisure class*. London: George Allen & Unwin.

Constant and Variable Capital

N. Okishio

Definition

In *Das Kapital* Marx defined Constant Capital as that part of capital advanced in the means of production; he defined Variable Capital as the

part of capital advanced in wages (Marx 1867, Vol. I, ch. 6). These definitions come from his concept of Value: he defined the value of commodities as the amount of labour directly and indirectly necessary to produce commodities (Vol. I, ch. 1). In other words, the value of commodities is the sum of C and N, where C is the value of the means of production necessary to produce them and N is the amount of labour used that is directly necessary to produce them. The value of the capital advanced in the means of production is equal to C.

However, the value of the capital advanced in wages is obviously not equal to N, because it is the value of the commodities which labourers can buy with their wages, and has no direct relationship with the amount of labour which they actually expend. Therefore, while the value of the part of capital that is advanced in the means of production is transferred to the value of the products without quantitative change, the value of the capital advanced in wages undergoes quantitative change in the process of transfer to the value of the products. This is the reason why Marx proposed the definitions of constant capital C and variable capital V.

The definition of constant capital and variable capital must not be confused with the definition of fixed capital and liquid capital. Fixed capital is a part of constant capital which is totally used in production process but transfers its value to products only partially. Liquid capital is a part of constant capital which is totally used up and transfers its whole value within one production process. So constant capital is composed of both fixed capital and liquid capital, and on the other hand liquid capital belongs partly to constant capital and partly to variable capital.

Marx introduced the concept 'value-composition of capital', μ, which is defined as the ratio of constant capital C to variable capital V:

$$\mu \equiv \frac{C}{V}. \tag{1}$$

Marx knew well that the value composition of capital reflects not only material characteristics of the process of production but also the social relationship between capitalists and labourers. In fact definition (1) can be rewritten as

$$\mu = \frac{C}{N} \cdot \frac{N}{V} \tag{2}$$

C/N reflects the character of the process of production and N/V reflects the class relationship between capitalists and labourers. C/N is the ratio of the amount of labour necessary to produce the means of production to the amount of labour directly bestowed, which is completely determined by the material condition in the process of production, while N/V is the ratio of the amount of labour which labourers actually expend to the amount of labour that is necessary in order to produce commodities which labourers can purchase with their wages. If labourers are forced to work longer with less wages, this ratio must rise.

Marx proposed to call the value-composition of capital, insofar as it is determined by the material condition of the process of production, 'the organic composition of capital'. More explicitly, 'The value-composition of capital, inasmuch as it is determined by, and reflects, its technical composition, is called the *organic* composition of capital' (*Capital*, Vol. III, ch. 8). However, as shown above, the value composition of capital is not determined by the material condition of the process of production alone. So it is better to introduce the ratio C/N in the place of the organic composition of capital, which is determined only by the material condition in the process of production. In order to avoid confusion, I call this ratio the 'organic composition of production'. This is the ratio of dead labour to living labour, which Marx himself frequently used in *Das Kapital*.

Variable Capital and Source of Profit

In contrast to Smith, Ricardo and others, Marx attached great importance to analysis to find the source of profit. He found that source in surplus labour, which is the excess of labour expended by labourers over the value of commodities which labourers can obtain with their wages (*Capital,*

vol. I, ch. 5). Using the notation introduced above, $N > V$ is the necessary condition for profit to exist. In order to illuminate this fact, he called capital advanced in wages *Variable Capital*. So the validity of this name depends on his analysis of the source of profit. How is it justified?

For simplicity we set up the simplest model which can reflect the fundamental characteristics of a capitalistic economy; these characteristics are the prevalence of commodity production, and the existence of class relationships between labourers and capitalists. There are only two kinds of commodities: the means of production (commodity 1) and consumption goods (commodity 2). In order to produce one unit of the ith commodity an amount of a_i unit of means of production and an amount of labour τ_t are necessary as input. Labourers are forced to work for T hours per day and earn the money wage rate w.

In order for profit to exist in both industries the following inequalities are necessary

$$p_1 > a_1 p_1 + \tau_1 w \tag{3}$$

$$p_2 > a_2 p_2 + \tau_2 w \tag{4}$$

where p_1 and p_2 denote the price of the means of production and consumption goods respectively. As labourers work for T hours a day at money wage w per hour, they can purchase an amount B of consumption goods.

$$B = \frac{wT}{p_2}, \quad B/T = R \tag{5}$$

where R is the real wage rate.

In the first volume of *Das Kapital,* Marx assumed that all commodities are exchanged at prices exactly proportionate to their unit value (equivalent exchange). Unit values of commodities are determined by the following equations

$$t_1 = a_1 t_1 + \tau_1 \tag{6}$$

$$t_2 = a_2 t_1 + \tau_2 \tag{7}$$

which assure unique and positive values, provided $a_1 < 1$ (Dmitriev 1898; May 1949–50; Okishio 1955a, b).

Under the assumption of equivalent exchange, we have

$$p_i = \lambda t_i \tag{8}$$

where λ is a constant which converts the dimension from hours to, say, dollars. Substituting (5) and (8) into (3) and (4) we get

$$t_1 > a_1 t_1 + \tau_1 \frac{B}{T} t_2 \tag{9}$$

$$t_2 > a_2 t_1 + \tau_2 \frac{B}{T} t_2 \tag{10}$$

By equations (6) and (7) and the above inequalities, we have

$$\tau_1 \left(1 - \frac{B}{T} t_2\right) > 0 \tag{11}$$

$$\tau_2 \left(1 - \frac{B}{T} t_2\right) > 0 \tag{12}$$

Consequently we arrive at the conclusion

$$T > B t_2. \tag{13}$$

This inequality implies the existence of surplus value, because surplus value is the excess of working hours T over the amount of labour necessary to produce commodities which labourers can receive with wages B. If the number of workers employed is n, then total expended labour is nT and variable capital measured in terms of value is $Bt_2 n$. So the inequality (13) can be rewritten as

$$N > V \tag{14}$$

This is the reason Marx called capital advanced in wages variable capital.

As shown above, Marx proved the theorem of the source of profit under the assumption of equivalent exchange. Though this is a clear-cut way to show the results, it has induced various critiques. Many critics have said that Marx's theorem would be right if all exchanges were equivalent

exchange, but that in reality exchanges are seldom equivalent so his theorem cannot be valid. In order to refute such a criticism we must prove the theorem without the assumption of equivalent exchange (see Okishio 1955a, 1955b, 1963, 1972; 1978; Morishima 1973). Mathematically, our task is to find necessary and sufficient conditions for inequalities (3), (4) and (5) to have non-negative solutions for p_1, p_2. From (3) we know easily that the condition

$$1 - a_1 > 0 \tag{15}$$

is necessary for p_1 to be positive. This condition ensures that the society will obtain net output. Next, substituting (5) into (3), and from (15) we have

$$\frac{p_1}{p_2} > \frac{\tau_1 B}{T(1 - a_1)}. \tag{16}$$

On the other hand, from (4) and (5) we get

$$\frac{p_1}{p_2} > \frac{T - \tau_2 B}{T a_2}. \tag{17}$$

We can easily get from (16) and (17)

$$\frac{a_2 \tau_1 B}{(1 - a_1)} < T - \tau_2 B. \tag{18}$$

Inequality (18) is rewritten as

$$T > B\left(\frac{a_2 \tau_1}{1 - a_1} + \tau_2\right). \tag{19}$$

By (19), (6) and (7) the above becomes

$$T > B t_2. \tag{20}$$

Thus we can arrive at Marx's result.

For later convenience we show another expression for the existence of surplus value. Dividing (3) and (4) by w, we get

$$\frac{p_1}{w} a_1 \frac{p_1}{w} + \tau_1 \tag{21}$$

$$\frac{p_2}{w} a_2 \frac{p_1}{w} + \tau_2 \tag{22}$$

By comparing (21) and (22), and (6) and (7), we get

$$\frac{p_i}{w} > t_i, \quad (i = 1, 2) \tag{23}$$

Equation (23) implies that if positive profit exists, then the price–wage ratio (the amount of commanded labour) is greater than the amount of value (necessary labour). In the famous controversy with Ricardo, Malthus pointed out this difference between labour commanded and labour embodied. Though he wrongly thought that this difference injured the validity of the labour theory of value, he had come near to the Marxian theory of the source of profit (see Malthus 1820, pp. 61–3, 120).

Condition (23) is rewritten as

$$1/t_i > w/p_i$$

This condition shows that if positive profit exists, then the productivity of labour ($1/t_i$) must be greater than the rate of real wages (w/p_i).

Organic Composition and Production Price

The concept of organic composition of capital plays an important role in Marx's analysis of prices.

The price of production (Ricardo's 'natural price') that gives every industry the equal rate of profit is determined by the following equations:

$$p_1 = (1 + r)(a_1 p_1 + \tau_1 w) \tag{24}$$

$$p_2 = (1 + r)(a_2 p_1 + \tau_2 w) \tag{25}$$

$$w = R p_2 \tag{26}$$

where r is the general (equal) rate of profit.

The first problem is to examine the relationship between

$$\frac{t_1}{t_2} \sim \frac{p_1}{p_2}.$$

If they are equal then we have equivalent exchange, if not we have non-equivalent exchange from the point of view of the labour theory of value. The values of the commodities are determined by (6) and (7). The ratio of the value of production-goods to consumption-goods t_1/t_2 is given as

$$\frac{t_1}{t_2} = \frac{\tau_1\left(\frac{a_1 t_1}{\tau_1} + 1\right)}{\tau_2\left(\frac{a_2 t_1}{\tau_2} + 1\right)} \tag{27}$$

The relative price of production-goods to consumption-goods determined by (24) and (25) is given as

$$\frac{t_1}{t_2} = \frac{\tau_1\left(\frac{a_1 t_1}{\tau_1} + w\right)}{\tau_2\left(\frac{a_2 p_1}{\tau_2} + w\right)}. \tag{28}$$

Comparing (27) with (28), we obtain

$$\frac{t_1}{t_2} - \frac{p_1}{p_2} = \frac{\tau_1}{\tau_2}\left[\frac{\frac{a_1 t_1}{\tau_1} + 1}{\frac{a_2 t_1}{\tau_2} + 1} - \frac{\frac{a_1 t_1}{\tau_1} + w}{\frac{a_2 p_1}{\tau_2} + w}\right]. \tag{29}$$

The expression in brackets on the RHS of (29) is given by

$$[\,] = (t_1 w - p_1)\left(\frac{a_1}{\tau_1} - \frac{a_2}{\tau_2}\right)A, \quad A > 0. \tag{30}$$

If profit is positive, from (23) $t_1 w - p_1$ is negative. So we can conclude

$$\frac{t_1}{t_2} \gtreqless \frac{p_1}{p_2} \Leftrightarrow \frac{a_1}{\tau_1} \lesseqgtr \frac{a_2}{\tau_2}. \tag{31}$$

The RHS of the above means the comparison of the organic composition of production and also the organic composition of capital, because as shown above the organic composition of production is $a_i t_1/\tau_i$ and the organic composition of capital is $a_i t_1/\tau_i\, R t_2$.

The second problem is to examine the influence of the change in real wage rate on the relative prices determined by (24), (25) and (26):

$$d\left(\frac{p_1}{p_2}\right) \Big/ \mathrm{d}R.$$

Denoting the relative price of production-goods to consumption-goods as p, from (24), (25) and (26) we obtain

$$f(p) \equiv a_2 p^2 + (\tau_2 R - a_1)p - \tau_1 R = 0. \tag{32}$$

Differentiating (32) with respect to R, we have

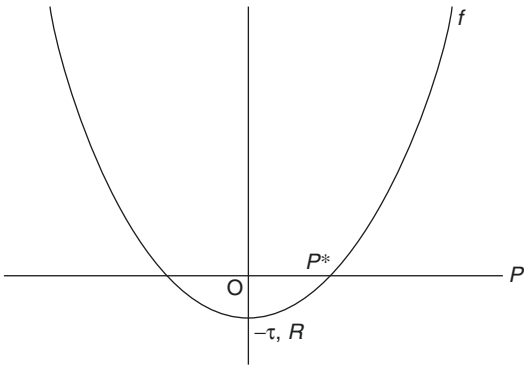
$$\frac{\mathrm{d}p}{\mathrm{d}R} = \frac{\tau_1 - \tau_{2p}}{2a_2 p + \tau_2 R - a_1}. \tag{33}$$

The denominator above is positive, because from (32)

$$\text{denominator} \times p = a_2 p^2 + \tau_1 R > 0.$$

We shall show that the sign of the numerator depends on the comparison between the organic composition of capital in both sectors.

The function $f(p)$ in (32) is drawn in Fig. 1. The meaningful solution of the equation (32) is given at p^*. Substituting t_1/t_2 into $f(p)$, we get



Constant and Variable Capital, Fig. 1

$$f\left(\frac{\tau_1}{\tau_2}\right) = \tau_1(a_2\tau_1 - a_2\tau_2).$$

Therefore if $a_2\tau_1 - a_1\tau_2 > 0$ then $f(\tau_1/\tau_2) > 0$, so considering the graph of $f(p)$ we know that $\tau_1/\tau_2 > p^*$. In the same way we can conclude that if $a_2\tau_1 - a_1\tau_2 \gtreqless 0$, then $\tau_1/\tau_2 \gtreqless p^*$. Consequently, from (3.10) we can conclude

$$d\left(\frac{p_1}{p_2}\right)/\mathrm{d}R \gtreqless 0 \Leftrightarrow \frac{a_1}{\tau_1} \lesseqgtr \frac{a_2}{\tau_2}.$$

This proposition is first established in Ricardo's *Principles* (1821, p. 43).

Organic Composition and the Rate of Profit

The concept of organic composition of capital plays an important role in Marx's analysis of the movement of the rate of profit (Fig. 1).

Marx defined the rate of profit as

$$r = \frac{S}{C+V}. \tag{34}$$

By (1), equation (34) is rewritten as

$$r = \frac{e}{\mu+1}, \quad e = S/V \tag{35}$$

where e is the rate of exploitation.

He asserted that if the organic composition of capital μ increases sufficiently then the rate of profit r must inevitably decrease. This is the famous 'law of the tendency for the rate of profit to fall' (*Capital,* Vol. III, Chap. 13).

Many people have criticized this theorem. They have said that if the rate of exploitation e increases sufficiently, r may increase in spite of the increase of μ. So r does not necessarily decrease, even if μ increases sufficiently (Robinson 1942; Sweezy 1942). Such a critique overlooks the logic of Marx's argument.

Marx stated:

> Since the mass of the employed living labour is continually on the decline as compared to the mass of materialized labour set in motion by it, i.e., to the productively consumed means of production, it follows that the portion of living labour, unpaid and congealed in surplus-value, must also be continually on the decrease compared to the amount of value represented by the invested total capital. Since the ratio of the mass of surplus-value to the value of the invested total capital forms the rate of profit, this rate must constantly fall (*Capital,* Vol. III, Chap. 13, p. 213).

Therefore Marx's true intention is to insist that if the organic composition of production $v = \mathrm{C/N}$ (the ratio of the mass of materialized labour to the mass of living labour) increases sufficiently, the rate of profit must fall.

This can be proved as follows (Okishio 1972). From (34) and (35), and

$$v = C/N \tag{36}$$

we have

$$\begin{aligned} r_{t+1} - r_t &= \frac{S_{t+1}}{C_{t+1}+V_{t+1}} - r_t \\ &= \frac{e_{t+1}}{v_{t+1}(1+e_{t+1})+1} - r_t \\ &= \frac{1}{v_{t+1}(1/e_{t+1}+1)+1/e_{t+1}} - r_t \end{aligned} \tag{37}$$

where suffixes t, $t+1$ denote periods.

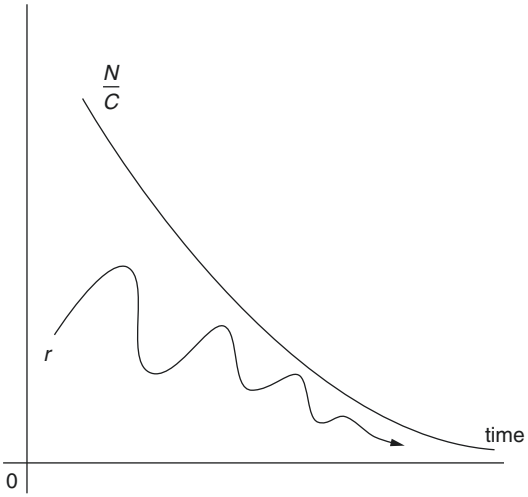
The RHS of (37) is an increasing function of e. If we take the limiting value as e tends to infinity, we have

$$r_{t+1} - t_t < \frac{1}{v_{t+1}} - r_t.$$

Therefore we conclude, if $v_{t+1} > 1/r_t$, then $r_{t+1} - r_t < 0$.

The above reasoning can be restated. The reciprocal of the organic composition of production sets an upper limit to the rate of profit, because

$$r = \frac{S}{C+V} < \frac{S+V}{C} = \frac{N}{C} \tag{38}$$



Constant and Variable Capital, Fig. 2

If this upper limit decreases sufficiently, the rate of profit must eventually decrease, as shown in Fig. 2.

In response to criticisms of this view we must say that as far as we accept Marx's assumption that the inverse of the organic composition (N/C) tends toward zero, Marx's conclusion inevitably follows.

So far we have defined the rate of profit as (34) and C, V, S are all measured in terms of labour value. However, the general rate of profit r must be determined by (24), (25) and (26). Can we derive the same conclusions for such a redefined r?

Eliminating p_1, p_2, w from (24), (25) and (26) we have

$$f(r,R)\equiv(1+r)^2R(a_1\tau_2-a_2\tau)-(1+r)(a_1+\tau_2R)+1=0 \tag{39}$$

Differentiating $f(r, R)$ we have

$$f_r\mathrm{d}r+f_R\mathrm{d}R=0 \tag{40}$$

where

$$f_r=2(1+r)R(a_1\tau_2-a_2\tau_1)-(a_1+\tau_2R)$$
$$f_R=(1+r)^2(a_1\tau_2-a_2\tau)-(1+r)\tau_2$$

Considering (39)

$$(1+r)f_r=(a_1+\tau_2R)(1+r)-2 \tag{41}$$

From (24), (25), (26), we know

$$1-(1+r)a_1>0 \quad 1-(1+r)\tau_2R>0 \tag{42}$$

From (41) $f_r<0$. f_R is rewritten as

$$f_R=(1+r)\{[(1+r)a_1-1]\tau_2-(1+r)a_2\tau_1\}$$

So by (42) $f_R<0$, from which $\mathrm{d}r/\mathrm{d}R<0$. As R goes to zero r tends to its upper limit, which is obtained from (39)

$$r_{\max}=\frac{1-a_1}{a_1}. \tag{43}$$

Since the value of the means of production is determined by (6), we have

$$\frac{1-a_1}{a_1}=\frac{(1-a_1)t_1}{a_1t_1}=\frac{\tau_1}{a_1t_1}=\frac{N_1}{C_1} \tag{44}$$

Thus the upper limit of the general rate of profit is given by the reciprocal of the organic composition of production in the means of production sector. Therefore if the organic composition in that sector rises sufficiently, the general rate of profit must fall.

Organic Composition and Unemployment

The concept of organic composition of capital plays an important role in Marx's analysis of the movement of employment (*Capital,* vol. I, ch. 23).

Marx assumed a rise in labour productivity to accompany the rise in the organic composition of production C/N. If C/N rises then from the definition of organic composition the amount of employment must decrease relative to constant capital.

However, how does the increase in the organic composition influence the absolute level of employment?

Many people thought that even if C/N rises sufficiently, still if constant capital C also increases then the absolute level of employment can also increase, though less than proportionately to constant capital (Oppenheimer 1903). But by reasoning similar to that used for 'the tendency of the rate of profit to fall', we can prove that if organic composition rises sufficiently, then the absolute level of employment must actually decrease.

The organic composition of production in the tth period v_t is defined as

$$v_t = \frac{C_t}{N_t}. \tag{45}$$

The accumulation of constant capital $\Delta C = C_{t+1} - C_t$ is financed from surplus value S.

$$C_{t+1} - C_t < S_t. \tag{46}$$

The surplus value S is a part of the amount of living labour which labourers expend

$$S_t < N_t. \tag{47}$$

By (45), we obtain,

$$\begin{aligned} N_{t+1} - N_t &= \frac{1}{v_{t+1}} C_{t+1} - \frac{1}{v_t} C_t \\ &= \frac{1}{v_{t+1}} (C_{t+1} - C_t) + C_t \left(\frac{1}{v_{t+1}} - \frac{1}{v_t} \right). \end{aligned}$$

From (46) and (47) we get

$$\begin{aligned} N_{t+1} - N_t &< \frac{1}{v_{t+1}} S_t + C_t \left(\frac{1}{v_{t+1}} - \frac{1}{v_t} \right) \\ &< \frac{N_t}{v_{t+1}} + C_t \left(\frac{1}{v_{t+1}} - \frac{1}{v_t} \right) \\ &= \frac{C_t}{v_{t+1} v_t} (1 + v_t - v_{t+1}) 0 \end{aligned}$$

we can say, if $(1 + v_t - v_{t+1}) < 0$ then $N_{t+1} - N_t < 0$. Therefore, if the organic composition of production in the $t + 1$th period, v_{t+1}, increases sufficiently so as to exceed $1 + v_t$, then the amount of employed labourer, N_{t+1} must inevitably become less than N_t, however high the rate of accumulation of capital may be (Okishio 1972). The rate of accumulation of capital $\Delta C/C$ itself is bounded by the reciprocal of the organic composition. From (46) and (47)

$$\frac{\Delta C}{C} < \frac{N}{C} = \frac{1}{v}$$

so that, because it is reasonable to assume that the growth rate of labour supply is non-negative, we can say that if the organic composition rises sufficiently the rate of unemployment inevitably rises. Though Marx did not state this explicitly, we think that this is what he wanted to say.

In analysing Marx's theorem on the movement of the rate of profit and employment, we have accepted his central assumption that the organic composition of production rises sufficiently over time. However, there arises the problem: under what conditions do capitalists choose techniques that have sufficiently high organic compositions of production?

Marx seemed to think that the rise in labour productivity and the rise in the organic composition are two aspects of the same thing. But these two do not always go together. Marx himself knew that if labour productivity in the means of production sector rises very high then even if technical composition rises, still the value composition may remain constant or decrease.

As to the capitalists' introduction of new techniques we have the following propositions:

(1) If the real wage rate remains constant and capitalists introduce new techniques which raise the rate remains of profit (calculated at the current prevailing prices and wage) then the new general rate of profit does not decrease, whatever the organic composition may be.
(2) If the real wage rate rises and capitalists adapt to this situation with the introduction of new techniques, then the new general rate of profit

does is higher than the one which would be expected if such a new technique were not introduced.

For the proofs of these propositions, see ▶ choice of technique and the rate of profit.

See Also

- ▶ Marxian Value Analysis
- ▶ Organic Composition of Capital
- ▶ Surplus Value

Bibliography

Dmitriev, V.K. 1898. The theory of value of David Ricardo. In *V.K. Dmitriev, economic essays on value, competition and utility*, ed. D.M. Nuti. Cambridge: Cambridge University Press, 1974.

Malthus, R. 1820. *Principles of political economy considered with a view to their practical application*, 1st ed. London: John Murray.

Marx, K. 1867–94. *Capital.* Translated from the third German edition by Samuel Moore and Edward Aveling, ed. Frederick Engels. New York: International Publishers, 1967.

May, K. 1949. The structure of classical theories. *Review of Economic Studies* 17(1): 60–69.

Morishima, M. 1973. *Marx's economics: A dual theory of value and growth*. Cambridge: Cambridge University Press.

Okishio, N. 1955a. Kachi to Kakaku (Value and production price). In *Keizaigaku Kenkyu Nempo* (The Annals of Economic Studies). Kobe University. No. 19.

Okishio, N. 1955a. Monopoly and the rates of profit. *Kobe University Economic Review* 1: 71–88.

Okishio, N. 1963. A mathematical note on Marxian theorems. *Weltwirtschaftliches Archiv* 91(pt. 2): 287–298.

Okishio, N. 1972. A formal proof of Marx's two theorems. *Kobe University Economic Review* 18: 1–6.

Okishio, N., et al. 1978. Three topics on Marxian fundamental theorems. *Kobe University Economic Review* 24: 1–18.

Oppenheimer, T. 1903. *Das Grundgesetz der Marxschen Gesellschaftslehre.* Book II, ch. 25. Berlin: Reimer.

Ricardo, D. 1821. On the principles of political economy and taxation. In *Works and correspondence of David Ricardo*, vol. 1, ed. P. Sraffa. Cambridge: Cambridge University Press, 1951–73.

Robinson, J. 1942. *An essay on Marxian economics*. London: Macmillan.

Sweezy, P.M. 1942. *The theory of capitalist development: Principles of Marxian political economy.* New York: Oxford University Press.

Constitutional Economics

James M. Buchanan

The term *Constitutional Economics* (Constitutional Political Economy) was introduced to define and to classify a distinct strand of research inquiry and related policy discourse in the 1970s and beyond. The subject matter is not new or novel, and it may be argued that 'constitutional economics' is more closely related to the work of Adam Smith and the classical economists than its modern 'non-constitutional' counterpart. Both areas of inquiry involve positive analysis that is ultimately aimed at contributing to the discussion of policy questions. The difference lies in the level of or setting for analysis which, in turn, implies communication with different audiences.

Orthodox economic analysis, whether this be interpreted in Marshallian or Walrasian terms, attempts to explain the choices of economic agents, their interactions one with another, and the results of these interactions, within the existing legal-institutional-constitutional structure of the polity. Normative considerations enter through the efficiency criteria of theoretical welfare economics, and policy options are evaluated in terms of these criteria. The policy analyst, building on the analysis, presents his results, whether explicitly or implicitly, to the political decision-makers, who then make some ultimate determination from among the available set. In this role the policy analyst directly, and the theorist indirectly, are necessarily advising governmental decision-makers, whoever these may be.

By both contrast and comparison, constitutional economic analysis attempts to explain the working properties of alternative sets of legal-institutional-constitutional rules that constrain

the choices and activities of economic and political agents, the rules that define the framework within which the ordinary choices of economic and political agents are made. In this sense, constitutional economics involves a 'higher' level of inquiry than orthodox economics; it must incorporate the results of the latter along with many less sophisticated subdisciplines. Normative considerations enter the analysis in a much more complex manner than through the artificially straightforward efficiency criteria. Alternative sets of rules must be evaluated in some sense analogously to ranking of policy options within a specified institutional structure, but the epistemological content of the 'efficiency' criteria becomes more exposed.

The constitutional economist, precisely because the subject matter is the analysis of alternate sets of rules, has nothing to offer by way of policy advice to political agents who act within defined rules. In this sense, constitutional economics is not appropriately included within 'policy science' at all. At another level, however, the whole exercise is aimed at offering guidance to those who participate in the discussion of constitutional change. In other words, constitutional economics offers a potential for normative advice to the member of the continuing constitutional convention, whereas orthodox economics offers a potential for advice to the practising politician. In a real sense, constitutional economics examines the *choice of constraints* as opposed to the *choice within constraints*, and as this terminology suggests, the disciplinary attention of economists has almost exclusively been placed on the second of these two problems.

A preliminary illustration of the distinction may be drawn from the economics of monetary policy. The constitutional economist is not directly concerned with determining whether monetary ease or monetary restrictiveness is required for furthering stabilization objectives in a particular setting. On the other hand, he is directly concerned with evaluating the properties of alternative monetary regime (e.g. rule-directed versus discretionary, fiat versus commodity standards). The ultimate objective of analysis is the choice among the institutions within which political agents act. The predicted behaviour of these agents is incorporated in the analysis of alternative sets of constraints.

Constitutional Economics and Classical Political Economy

As suggested, Constitutional Economics is related to classical political economy and it may be considered to be an important component of a more general revival of the classical emphasis, and particularly as represented in the works of Adam Smith. (The closely related complementary components are discussed briefly in section "The New Political Economy".) One obvious aim of the classical political economists was to offer an explanation and an understanding of how markets operate without detailed political direction. In this respect, orthodox neoclassical economics follows directly in the classical tradition. But the basic classical analysis of the working of markets was only a necessary step toward the more comprehensive purpose of the whole exercise, which was that of demonstrating that, precisely because markets function with tolerable efficiency independently of political direction, a powerful normative argument for constitutional structure exists. That is to say, Adam Smith was engaged directly in comparing alternative institutional structures, alternative sets of constraints within which economic agents make choices. In this comparative analysis, he found it essential to model the working properties of a non-politicized economy, which did not exist in reality, as well as the working properties of a highly politicized mercantilist economy, which could be directly observed.

There is no need here to enter the lists on either side of the 'ideas have consequences' debate. We know that the economy of Great Britain was effectively de-politicized in the late 18th and early 19th centuries, and from the analysis of Smith and his classical fellow travellers there emerged both positive understanding of economic process and philosophical argument for a particular regime. The normative argument for laissez faire was, perhaps inevitably, intermingled with the positive analysis of interaction within a

particular structure of constraints, essentially those that describe the minimal, protective, or night-watchman state. Economics, as a social science, emerged, but in the process attention was diverted from the institutional structure. Even the predicted normative reaction against the overly zealous extension of the laissez faire economics argument was couched in 'market failure' terms, rather than in the Smithian context of institutional comparison. The early socialist critique of market order, both in its Marxist and non-Marxist variants, was almost exclusively negative in that it elaborated putative failures of markets within an unexamined set of legal-political rules while it neglected analysis of the alternative rules that any correction of the alleged failures might require. Only with the debates on socialist calculation in the decades prior to World War II did the issues of comparative structure come to be examined.

It was only in the half-century after these debates that political economy, inclusively defined, returned, in fits and starts, to its classical tradition. Given the legal order of the protective state (the protection of property and the enforcement of contracts), we now know that under some conditions 'markets fail' when evaluated against idealized criteria, whether these be 'efficiency', 'justice', or other abstract norms. We also know that 'politics fails' when evaluated by the same criteria. Any positive analysis that purports to be of use in an ultimate normative judgment must reflect an informed comparison of the working properties of alternative sets of rules or constraints. This analysis is the domain of Constitutional Economics.

Constitutional Economics and Social Philosophy

Classical political economy emerged from moral philosophy, and its propounders considered their efforts to fall naturally within the limits of philosophical discourse. As a modern embodiment, Constitutional Economics is similarly located, regardless of disciplinary fragmentation. How can persons live together in liberty, peace and prosperity? This central question of social philosophy requires continuing contributions from many specialists in inquiry, surely including those of the constitutional economists. By their focus directly on the ultimate selection of a set of constraining rules within which ordinary social interaction takes place, constitutional economists remove themselves at least one stage further from the false position of 'social engineer' than their counterparts in orthodox economics. Precisely because there is no apparently simple evaluative criterion analogous to 'allocative efficiency' at hand, the constitutional economist is less tempted to array alternatives as if an unexamined criterion commands universal assent. The artificial abstraction of 'social utility' is likely to be less appealing to those who concentrate on choices among constraints than to those who examine choices within constraints.

If, however, there is no maximand, how can ultimate normative consequence emerge? In this respect, one contribution lies at the level of positive analysis rather than in a too-hasty leap into normative evaluation. Classical political economy contains the important principle of spontaneous coordination, the great discovery of the 18th century. This principle states that, within the legal umbrella of the minimal state and given certain conditions, the market 'works'. Even if in the principle's modern embellishment we must add 'warts and all', we still have come a long way toward a more comprehensive understanding of the alternatives for social order. To economics the extent that his efforts expand the public understanding of this principle, in application to all institutional settings, the constitutional economist remains under less apparent compulsion to advance his own privately preferred 'solutions' to the ultimate choice among regimes.

The New Political Economy

Care should be taken not to claim too much for Constitutional Economics, especially if a narrow definition is used. As noted earlier, this research programme, by designation, emerged in the 1970s to describe efforts at analysing the effects of

alternative sets of rules, as opposed to analyses of choices made within existing and unexamined structures. In a more comprehensive overview of developments after World War II, Constitutional Economics takes its place among an intersecting set of several research programmes, all of which have roots in classical political economy. Critical emphases differ as among the separate programmes, but each reflects efforts to move beyond the relatively narrow confines of orthodox neoclassical economics.

In continental Europe, the whole set of subdisciplines is included under the rubric 'The New Political Economy'. Within this set we can place (1) Public Choice, from which Constitutional Economics emerged; (2) Economics of Property Rights; (3) Law and Economics or Economic Analysis of Law; (4) Political Economy of Regulation; (5) the New Institutional Economics, and (6) the new Economic History. Defined imperialistically, Constitutional Economics would parallel the inclusive term and embrace all of these programmes, since some attention is drawn in each case to the legal-political constraints within which economic and political agents choose. Differences can be identified, however, and it may be useful to summarize some of these here, even if detailed discussion of the other research programmes cannot be attempted.

Public Choice, in its non-constitutional aspects of inquiry, concentrates attention on analyses of alternative political choice structures and on behaviour within those structures. Its focus is on predictive models of political interactions, and is a preliminary but necessary stage in the more general constitutional inquiry. The economics of property rights, law and economics, and the political economy of regulation remain somewhat closer to orthodox economic theory than Constitutional Economics or Public Choice. The standard efficiency norm remains central to these subdisciplines, both as an explanatory benchmark and as normative ideal. The new institutional economics is directed more toward the interactions within particular institutional forms rather than toward the comprehensive structure of political rules (Furubotn and Richter 1980; Frey 1984). Some elements of the new economic history closely parallel Constitutional Economics, with, of course, an historical rather than a comparative emphasis (North and Thomas 1973).

Presuppositions

Constitutional Economics, along with the related research programmes mentioned above, shares a central methodological presupposition with both its precursor, classical political economy, and its counterpart in modern neoclassical microeconomics. Only individuals choose and act. Collectivities, as such, neither choose nor act and analysis that proceeds as if they do is not within the accepted scientific canon. Social aggregates are considered only as the results of choices made and actions taken by individuals. The emphasis on explaining non-intended aggregative results of interaction has carried through since the early insights of the Scottish moral philosophers. An aggregative result that is observed but which cannot, somehow, be factored down and explained by the choices of individuals stands as a challenge to the scholar rather than as some demonstration of non-individualistic organic unity.

Methodological individualism, as summarized above, is almost universally accepted by economists who work within mainstream, or non-Marxian, traditions. A philosophical complement of this position that assumes a central role in Constitutional Economics is much less widely accepted and is often explicitly rejected. A distinction must be drawn between the methodological individualism that builds on individual choice as the basic unit of analysis and a second presupposition that locates the ultimate sources of value exclusively in individuals.

The first of these presuppositions without the second leaves relatively little scope for the derivation of constitutional structures from individual preferences. There is no conceptual normative bridge between those interests and values that individuals might want to promote and those non-individualistic values that are presumed to serve as ultimate normative criteria. The whole constitutional exercise loses most if not all of its *raison d'être* in such a setting. If the ultimate

values which are to be called upon to inform the choices among institutions are non-individualistic, then there is, at best, only an instrumental argument for using individually expressed preferences in the process of discovering those values.

On the other hand, if the second presupposition concerning the location of the ultimate sources of value is accepted, there is no *other* means of deriving a 'logic of rules' than that of utilizing individually expressed interests. At base, the second presupposition implies democracy in governance, along with the accompanying precept that this structure of decision-making only takes on normative legitimacy with the prefix 'constitutional' appended to it.

Wicksell as Precursor

The single most important precursor to Constitutional Economics in its modern variant is Knut Wicksell, who was individualist in both of the senses discussed above. In his basic work on fiscal theory (*Finanztheoretische Untersuchungen*, 1896), Wicksell called attention to the significance of the rules within which choices are made by political agents, and he recognized that efforts at reform must be directed toward changes in the rules for making decisions rather than toward modifying expected results through influence on the behaviour of the actors.

In order to take these steps, Wicksell needed some criterion by which the possible efficacy of a proposed change in rules could be judged. He introduced the now-familiar unanimity or consensus test, which is carried over into Constitutional Economics and also allows the whole research programme to be related closely to the contractarian tradition in political philosophy. The relationship between the Wicksellian and the Paretian criteria is also worthy of note. If only individual evaluations are to count, and if the only source of information about such evaluations is the revealed choice behaviour of individuals themselves, then no change could be assessed to be 'efficient' until and unless some means could be worked out so as to bring all persons (and groups) into agreement. If no such scheme can be arranged, the observing political economist remains silent. The Wicksellian contribution allowed the modern economist to bring the comparative analysis of rules or institutions within a methodological framework that utilizes and builds on the efficiency criterion, which, when interpreted as indicated, does not require departure from either of the individualistic presuppositions previously discussed.

Homo Economics in Constitutional Choice

Constitutional Economics, as distinct from the complementary research programme on political constitutions that are within the boundaries of law, political science, sociology and other disciplines, goes beyond the logical presuppositions of individualism to incorporate nontautological models of individual utility maximization. *Homo economicus* takes a central role in comparative institutional inquiry. Individuals are assumed to seek their own interests, which are defined so as to retain operational content.

Two quite different arguments can be made in support of this postulate in Constitutional Economics. The first is based simply on methodological consistency. To the extent that individuals are modelled as utility maximizers as they participate in market relationships, there would seem to be no basis for postulating a shift in motivation as they behave within non-market constraints. There is at least a strong presumption that individuals do not undergo character transformation when they shift from roles as buyers or sellers in the market-place to roles as voters, taxpayers, beneficiaries, politicians, or bureaucrats in the political process. A more sophisticated reason for postulating consistency in behaviour lies in the usefulness of the model for the whole exercise of institutional comparison. If the purpose is to compare the effects of alternative sets of constraints, some presumption of behavioural consistency over the alternatives is necessary in order to identify those differences in results that are attributable to the differences in constraints.

A second argument for introducing *homo economicus* in Constitutional Economics is both more complex and more important. It is also the source of confusion because it is necessary to distinguish carefully between the use of *homo economicus* in predictive social science, specifically in positive Public Choice and in neoclassical economics, and in Constitutional Economics. There is an argument for using the construction in the latter, even if there are demonstrated empirical limits on the explanatory power of the model in the former.

The argument is implicit in the work of the classical economists. It was stated as a methodological principle by both David Hume and J.S. Mill:

> In constraining any system of government, and fixing the several checks and controls of the constitution, each man ought to be supposed a knave, and to have no other end, in all his actions, than private interest. (Hume [1741], 1963, pp. 117–18)
>
> The very principle of constitutional government requires it to be assumed that political power will be abused to promote the particular purposes of the holder; not because it is always so, but because such is the natural tendency of things, to guard against which is the special use of free institutions. (Mill [1861], 1977, p. 505)

The ultimate purpose of analysing alternative sets of rules is to inform the choice among these sets. The predicted operating properties of each alternative must be examined, and these properties will reflect the embodied models of individual behaviour within the defined constraints. Behavioural departures from the presumptive models used in deriving the operating properties will, of course, be expected. But the costs of errors may not be symmetrically distributed around the single best predictive model. The predicted differential loss from behavioural departures from a model that involves 'optimistic' motivational assumptions may be much larger than the predicted differential gain if the model is shown to be an accurate predictor. Hence, comparative evaluation of an institution based on an altruistic model of behaviour should take into account the possible non-linearity in the loss function that describes departures from the best estimates. (In legal practice, formal contracts include protections against worst-case behaviour patterns.) In constitutional choice, therefore, there is an argument for incorporating models of individual behaviour that presume more narrowly defined self-interest than any empirical record may warrant (Brennan and Buchanan 1985).

Applications

Applications of Constitutional Economics, as a research programme, have emerged in several settings. First, consider taxation. Post-Marshallian economic theory, either in its partial or general equilibrium model, was often applied to tax incidence. Analysis was directed toward predicting the effects of an exogenously imposed tax on the private economizing behaviour of persons in their varying capacities as demanders and suppliers of goods and services in the market-place. Building on this base of positive analysis, normative welfare economics allows a ranking among alternative equi-revenue tax instruments in terms of the Paretian standard. In both the positive and normative aspects, neoclassical tax theory embodies the presumption that taxes, as such, are exogenous to the choice process.

The major contribution of modern Public Choice, as a subdiscipline in its own right, has been that of endogenizing political decision-making. In its direct emphasis, public choice theory examines the political decision rules that exist with a view toward making some predictions about just what sort of tax institutions or tax instruments will emerge. Constitutional Economics, as an extended research programme that emerges from Public Choice, goes a step further and uses the inputs from both neoclassical economics and public choice theory to analyse how alternative political rules might generate differing tax rules.

The relevant constitutional choice may be that of granting government authority to levy taxes on Tax Base A or Tax Base B. Suppose that under the neoclassical equi-revenue assumption, analysis demonstrates that the taxing of A generates a lower excess burden than the taxing of B. Analysis of the political choice process may

demonstrate, however, that government, if given the authority to tax A, will tend to levy a tax that will generate *more* revenue than would be forthcoming under an authority to tax B. The equi-revenue alternatives may not be effective political alternatives under any plausibly acceptable modelling of the behaviour of political agents. Once this simple point is recognized, the normative significance of the neoclassical ranking of tax instruments is reduced. Discussion shifts necessarily to the level of interaction between political decision structures and fiscal institutions.

A second application of Constitutional Economics is found in the post-Keynesian discussion of budgetary policy. The Keynesian advocacy of the use of governmental budgets to accomplish macroeconomic objectives was based on a neglect of the political decision structure. The proclivity of democratic governments to prefer spending over taxing, and hence to bias budgets toward deficit, is readily explained in elementary public choice theory (Buchanan and Wagner 1977). This essential step in public choice reasoning leads naturally to inquiry into the relationships between the constraints that may be placed on political choice and predicted patterns of budgetary outcomes. Out of this intensely practical, and important, application of Constitutional Economics emerged the intellectual bases for the normative argument that, in the post-Keynesian era when moral constraints on political agents have lost much of their previous effectiveness, formal rules limiting deficit financing may be required to insure responsible fiscal decisions. In the modern setting, such rules would limit spending rates. But it is perhaps worth noting that, in the political environment of Sweden in the 1890s, Wicksell advanced analytically similar proposals for reform in the expectation that, if the suggested reforms should be implemented, public sector outlay would increase.

The analysis of alternative rules for 'the transfer constitution' represents a third application of constitutional economics. With the 1971 publication of John Rawls's *A Theory of Justice*, renewed attention came to be placed on principles of distributive justice. Although explicitly pre-constitutional, Rawls's work has a close relationship with the efforts to derive criteria for political and economic rules of social interaction. Economists, as well as other social scientists and social philosophers, have come increasingly to recognize that the untrammelled interplay of interest-group politics is unlikely to further objectives for distributive justice. Analysis of how this politics operates in the making of fiscal transfers suggests that principled adjustments in the post-tax, post-transfer distribution of values is only likely to be achieved if the institutional rules severely restrict the profitability of investment in attempts to subvert the transfer process.

Further applications include the regulatory constitutions, along with the organization of public enterprises. In its inclusive definition, Constitutional Economics becomes the analytical route through which institutional relevance is reintroduced into a sometimes sterile social science. In its less inclusive definition, Constitutional Economics, along with its related and complementary research programmes, restores 'political' to 'economy', thereby bringing a coherence that was absent during the long hiatus during which 'economics' made putative claims to independent status.

See Also

- ▶ Black, Duncan (1908–1991)
- ▶ Law and Economics
- ▶ Public Choice
- ▶ Social Choice
- ▶ Voting

References

Brennan, G., and J.M. Buchanan. 1980. *The power to tax: Analytical foundations of the fiscal constitution*. Cambridge: Cambridge University Press.

Brennan, G., and J.M. Buchanan. 1985. *The reason of rules: Constitutional political economy*. Cambridge: Cambridge University Press.

Buchanan, J.M. 1974. *The limits of liberty: Between anarchy and leviathan*. Chicago: University of Chicago Press.

Buchanan, J.M., and G. Tullock. 1962. *The calculus of consent: Logical foundations of constitutional democracy*. Ann Arbor: University of Michigan Press.

Buchanan, J.M., and R.E. Wagner. 1977. *Democracy in deficit: The political legacy of Lord Keynes*. New York: Academic.
Frey, B. 1984. A new view of economics: Comparative analysis of institutions. *Scelte Pubbliche* 1: 17–28.
Furubotn, E.G, Richter, R. (eds). 1980. *The new institutional economics – A symposium*. Zeitschrift für die gesamte Staatswissenschaft, 140.
Hayek, F.A. 1973–1979. *Law, legislation, and liberty*, 3 vols. Chicago: University of Chicago Press.
Hume, David. 1741. On the interdependency of parliament. In *Essays, moral, political and literary*. London: Oxford University Press, 1963.
McKenzie, R. 1982. *Bound to be free*. Palo Alto: Hoover Press.
McKenzie, R. (ed.). 1984. *Constitutional economics*. Lexington: Lexington Books.
Mill, J.S. 1861. Considerations on representative government. In *Essays on politics and society*, vol. XIX of Collected works of J.S. Mill. Toronto: University of Toronto Press, 1977.
North, D.C., and R.P. Thomas. 1973. *The rise of the western world: A new economic history*. Cambridge: Cambridge University Press.
Rawls, J. 1971. *A theory of justice*. Cambridge, MA: Harvard University Press.
Wicksell, K. 1896. *Finanztheoretische Untersuchungen*. Jena: Gustav Fischer. Central portions of this work were published in English translation as 'A new principle of just taxation'. In *Classics in the theory of public finance*, ed. R.A. Musgrave and A.T. Peacock. London: Macmillan, 1959.

Constitutions, Economic Approach to

Dennis C. Mueller

Abstract

The economic approach to constitutions applies the methodology of economics to the study of constitutions. This entry reviews the normative literature on constitutions, which assumes a two-stage collective decision process, and the positive literature that examines the decisions made by constitutional conventions and their economic consequences.

Keywords

Beard, C; Bentham, J; Buchanan, J; Budget deficits; Collective choice; Constitutionalism normative vs positive;; Corruption; First amendment; Harsanyi, J; Landes, W; Negative externalities; Party systems; Philadelphia convention; Posner, R; Rawls, J; Rent seeking; Social contract theory; Social welfare function; Tullock, G

JEL Classification

K1

The economic approach to constitutions applies the methodology of economics to the study of constitutions, just as public choice applies this methodology to the full range of topics of political science.

The economic approach to constitutions began with *The Calculus of Consent* by James Buchanan and Gordon Tullock (1962, hereafter B&T). Theirs was largely a *normative* analysis of what ought to go into a constitution. Their main findings and the literature that grew out of their work are reviewed first, after which the *positive* stream of the literature is discussed.

Normative Research on Constitutions

Arguably the most important contribution of *The Calculus* was to view democracy as a two-stage process. In stage one, institutions to make future collective decisions are placed into the constitution. In stage two, collective decisions are made using these rules. The long-run nature of the choices at the first stage creates considerable uncertainty about the consequences of different voting rules. This uncertainty makes unanimous agreement on the rules of the political game likely, even though individuals would disagree in stage two about the outcomes of the game. This unanimity at the constitutional stage provides the normative underpinning for the constitution (B&T, p. 7). Harsanyi (1955) also used uncertainty over future positions to produce unanimity and to provide a normative argument for a Benthamite social welfare function (SWF), as did Rawls (1971) in his ethical theory of a social contract. Mueller (1973) discussed conditions

under which a B&T constitution maximizes a Harsanyian SWF.

Another innovation in *The Calculus* was to introduce the *external costs of collective decisions* (B&T, pp. 63–8). When a collective choice is made without the consent of all members of the community, the decision can make some members worse-off. The votes of those favouring the decision thus impose a negative externality on those opposing it. The smaller the majority required to pass an issue, the more likely it is that an individual is on the losing side. However, the amount of time required to make a collective decision is also likely to increase with the required majority. The optimal majority minimizes the sum of collective decisions' external and decision-making costs.

There is nothing in B&T's costs-minimization-approach that implies that the optimal majority is likely to be a simple majority, and thus their approach does not account for this rule's ubiquitous use. The approach does imply the widespread use of the simple majority rule, if one of the two cost curves – most plausibly decision-making costs – has a sharp discontinuity at 50% (Mueller 2003, pp. 76–8).

Rae (1969) used the two-stage approach to provide a completely different normative justification for the simple majority rule. At the constitutional stage, each individual is uncertain of whether he will favour x or $\sim x$ in future votes on these binary issues. The expected gain if an individual favours x and x wins equals the expected loss if x wins and the individual favours $\sim x$. Rae further assumed that the probability of favouring x equals the probability of favouring $\sim x$. An egoist chooses the voting rule that minimizes the probability that she favours x in the future and $\sim x$ is imposed, or that she favours $\sim x$ and x is imposed. The simple majority is the only rule satisfying this condition. (For additional discussion and references see, Rae and Schickler 1997.)

Mueller (2001) generalized the two-stage approach to show that the optimal majority for binary choices depends on the relative payoffs from the two issues. (Riley 2001, presents a game theoretic analysis of a two-stage constitutional process.) As the loss to those favouring x rises relative to the gain to those favouring $\sim x$, higher required majorities become optimal to implement $\sim x$, with unanimity being optimal when the asymmetry in payoffs is very large. Mueller (1991, 1996, ch. 14) employed this analysis to explain why placing rights to act into a constitution would maximize the expected utilities of those writing it.

C

Positive Research on Constitutions

The positive literature of constitutions falls into two categories: studies of constitutional conventions and of the consequences of constitutions. The second category is obviously very large, and so I provide only the flavour of this type of work.

Charles Beard's work (1913) might well be regarded as the first *economic* analysis of the Philadelphia Convention. Beard stressed the self-interest of the participants, and claimed that the final product reflected the interests of the landowning aristocracy. In an equally cynical analysis, Landes and Posner (1975, p. 893) claimed that the First Amendment was a result of pressure from 'publishers, journalists, pamphleteers, and others who derive pecuniary and nonpecuniary income from publication and advocacy of various sorts'. Case studies of constitutional conventions confirm the importance of the self-interest of the participants in determining the constitution's content. For example, representatives from small parties favour rules that produce proportional representation and low percentage thresholds for taking seats in the parliament. Representatives from large parties favour the reverse. If delegates are selected geographically, the constitution protects geographic interests. (For further discussion and references to the literature, see Elster 1991, and Mueller 1996, ch. 21). Econometric analyses confirm these findings. McGuire and Ohlsfeldt (1986) and McGuire (1988) concluded that the votes of delegates to the Philadelphia convention reflected both their personal interests and those of their constituencies. Eavey and Miller (1989) reached the same conclusion from the voting patterns of those who ratified the Pennsylvania and Maryland constitutions.

A key decision facing any constitutional convention is whether to design institutions that will

produce a two-party system or a multiparty system. In practice, this choice appears to rest upon the number of representatives elected from each electoral district (Taagepera and Shugart 1989; Lijphart 1990; Mueller 1996, chs. 8–10). Recent theoretical and empirical work by Persson and Tabellini (1999, 2000, 2003, 2004a, b) and Persson et al. (2000) demonstrates the economic importance of this choice. They find more rent seeking, more corruption, more redistribution and larger deficits in multiparty systems. Presidential systems lead to smaller governmental sectors because they generally contain stronger checks and balances than parliamentary systems. (For a review and references to other contributions, see Persson and Tabellini 2004a.)

Conclusions

There are two kinds of people in the world: those who believe that constitutions matter and those who do not. The contributors to the literature reviewed here fall into the former category. Their work helps illustrate *why and in what way* constitutions matter, and further illustrates the fruitfulness of undertaking an economic approach to the study of constitutions.

See Also

- ▶ Buchanan, James M. (Born 1919)
- ▶ Collective Rationality

Bibliography

Beard, C. 1913. *An economic interpretation of the constitution of the United States*. New York: Macmillan, 1941.

Buchanan, J., and G. Tullock. 1962. *The calculus of consent*. Ann Arbor: University of Michigan Press.

Eavey, C., and G. Miller. 1989. Constitutional conflict in state and nation. In *The federalist papers and the new institutionalism*, ed. B. Grofman and D. Wittman. New York: Agathon Press.

Elster, J. 1991. *Arguing and bargaining in two constituent assemblies*. Mimeo. Storrs Lectures, Yale Law School.

Harsanyi, J. 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economics* 63: 309–321.

Landes, W., and R. Posner. 1975. The independent judiciary in an interest-group perspective. *Journal of Law and Economics* 18: 875–901.

Lijphart, A. 1990. The political consequences of electoral laws, 1945–85. *American Political Science Review* 84: 481–496.

McGuire, R. 1988. Constitution making: A rational choice model of the Federal Convention of 1787. *American Journal of Political Science* 32: 483–522.

McGuire, R., and R. Ohlsfeldt. 1986. An economic model of voting behavior over specific issues at the constitutional convention of 1787. *Journal of Economic History* 46: 79–111.

Mueller, D. 1973. Constitutional democracy and social welfare. *Quarterly Journal of Economics* 87: 60–80.

Mueller, D. 1991. Constitutional rights. *Journal of Law, Economics, and Organization* 7: 313–333.

Mueller, D. 1996. *Constitutional democracy*. Oxford: Oxford University Press.

Mueller, D. 2001. The importance of uncertainty in a two-stage theory of constitutions. *Public Choice* 108: 223–258.

Mueller, D. 2003. *Public choice III*. Cambridge: Cambridge University Press.

Persson, T., and G. Tabellini. 1999. The size and scope of government: Comparative politics with rational politicians. *European Economic Review* 43: 699–735.

Persson, T., and G. Tabellini. 2000. *Political economics – Explaining economic policy*. Cambridge, MA: MIT Press.

Persson, T., and G. Tabellini. 2003. *Economic effects of constitutions*. Cambridge, MA: MIT Press.

Persson, T., and G. Tabellini. 2004a. Constitutions and economic policy. *Journal of Economic Perspectives* 18: 75–98.

Persson, T., and G. Tabellini. 2004b. Constitutional rules and fiscal policy outcomes. *American Economic Review* 94: 25–45.

Persson, T., G. Roland, and G. Tabellini. 2000. Comparative politics and public finance. *Journal of Political Economy* 108: 1121–1161.

Rae, D. 1969. Decision-rules and individual values in constitutional choice. *American Political Science Review* 63: 40–56.

Rae, D., and E. Schickler. 1997. Majority rule. In *Perspectives on public choice*, ed. D.C. Mueller. Cambridge, MA: Cambridge University Press.

Rawls, J. 1971. *A theory of justice*. Cambridge, MA: Belknap.

Riley, J. 2001. Constitutional democracy as a two-stage game. In *Constitutional culture and democratic rule*, ed. J. Ferejohn, J. Rakove, and J. Riley. Cambridge: Cambridge University Press.

Taagepera, R., and M. Shugart. 1989. *Seats and votes*. New Haven: Yale University Press.

Consumer Durables

John Muellbauer

Applied work on the demand for durable goods has usually analysed two kinds of data. The first is time-series data on purchases, aggregated over consumers and typically with different kinds of durables aggregated into one or two groups. The second is cross-section data on the ownership of different kinds of durables. There has been a corresponding specialization in economic theory with that appropriate to the first type of data neglecting the issues of discreteness of ownership emphasized in the second and instead focusing on the dynamics of investment, expectations and adjustment costs, these being neglected in the theory of discrete choice at the level of individual households. The discussion below is in this tradition. In the first part, the focus is on the dynamics of purchases and in the second on the microeconomics of discrete choice.

A good which is durable yields a flow of services into the future. Whether other issues arise, the analysis of the demand for durables must take into account the distinction between stocks of goods and flows of services and the intertemporal character of the decision to purchase or own a durable good. The simplest coherent model which captures these two essential features was first exposited by Cramer (1957) though somewhat analogous analyses of the demand for investment goods had been in the literature for some time (Fisher 1930).

The assumptions made in this model are the following. Consumers maximize utility through time. They can lend and borrow at the same interest rate. They can instantly buy or sell a durable good at the same price. There are no psychic adjustment costs or habits associated with purchasing or owning a durable good. The service flow from owning a durable is proportional to the stock. Deterioration of the service flow through time is geometric. New vintages of durables are exactly the same as the old, when converted into efficiency units. Durable goods are perfectly divisible and no discreteness issues arise: thus they need not be owned in integer amounts and the question of scrapping never arises though durables can be traded on the second-hand market. Stating the assumptions this baldly when it is plain how counter-factual many of them are serves to anticipate some of the discussion below.

On these 'neoclassical' assumptions, the consumer maximizes

$$u_t = V\left(q_t, S_t, q_{t+1}, S_{t+1}, \ldots, q_T, S_T, A_T\right) \qquad (1)$$

subject to a sequence of period to period budget constraints of the form, for example, at t,

$$A_t = A_{t+1}(1 + r_t) + y_t - q_t - \left(P_t^D / P_t\right)(S_t - (1 - \delta)S_{t-1}) \qquad (2)$$

where q is the flow of non-durable consumption, S is the stock of durables and A is the stock of real financial assets defined as the nominal stock deflated by the price index for non-durable goods. Its presence in (1) reflects the bequest motive. The budget constraint (2) which looks slightly formidable just expresses the fact that the change in financial assets equals financial saving, i.e. income minus expenditure. The change in real financial assets is $A_t - A_{t-1}$ and income is $r_t A_{t-1} + yt$, where r is the real rate of return and y is real non-property income. Note that the real rate of return includes any capital gains or losses so that property income $r_t A_{t-1}$ is an economist's rather than a national income accountant's measure. Expenditure in real terms, i.e. in terms of the flow of non-durable consumption, consists of q_t and money expenditure on durables deflated by the price index for non-durables p_t. Money expenditure on durables is $p_t^D[s_t - (1 - \delta)s_{t-1}]$, where p^D is the price index for durables and where the stock of durables $S_t = q_t^D + (1 - \delta)s_{t-1}$ where q_t^D is the flow of purchases and δ is the rate of deterioration of durables so that $(1 - \delta)S_{t-1}$ is the amount of durables owned at $t - 1$ that survives into period t. The generalization of (1) and (2) when q and S are vectors is easy.

It is conventional in intertemporal consumer theory under point expectations to convert the

period to period budget constraints into the present value form by eliminating financial assets from the sequence of constraints (2) for t, t+1, $t + 2$, ... Life-cycle wealth is defined as the value of initial durables plus financial assets plus human wealth (the present value of current and expected non-property income):

$$W_t \equiv p_t^D/p_t(1-\delta)S_{t-1} + A_{t-1}(1+r_t) + \sum_{j=t}^{T} \rho_j y_j. \tag{3}$$

where the real discount factor ρ_j is defined by

$$\rho_j = \prod_{i=t+1}^{j} (1+r_i)^{-1}, \quad t+1 \leqq j \leqq T$$
$$\rho_j = 1. \tag{4}$$

The budget constraint for the decision variables q_t , S_t , q_{t+1} , S_{t+1} etc. correspondingly is:

$$W_t = \sum_{j=1}^{T} \rho_j q_j + (p_t^D/p_t)S_t + \sum_{j=t+1}^{T} \rho_j (p_t^D/p_j)\left[S_j - (1-\delta)S_{j-1}\right]. \tag{5}$$

The effective relative price associated with S_t can then be written in the 'user cost' form

$$\begin{aligned} \Pi_t &= (p_t^D/p_t) \\ &\times \left\{1 - (1-\delta)\left[1 + \frac{\Delta(p_{t+1}^D/p_{t+1})}{p_{t+1}^D/p_{t+1}}\right]/1 + r_{t+1}\right\} \\ &\times \approx p_t^D/p_t[\delta - \Delta \ln](p_{t+1}^D/p_{t+1}) \\ &+ r_{t+1} \end{aligned} \tag{6}$$

which is the approximation familiar from Jorgenson's (1963) neoclassical theory of investment. Thus (5) becomes

$$W_t = \sum_{j=t}^{T} \rho_j q_j + \sum_{j=t}^{T} \rho_j \Pi_j S_j. \tag{7}$$

Maximizing (1) subject to (7) gives demand functions

$$\begin{aligned} q_t &= g(\rho, \rho\Pi, W_t) \\ S_t &= g^D(\rho, \rho\Pi, W_t). \end{aligned} \tag{8}$$

where ρ , $p\prod$ are the vectors (ρ_j) , $(\rho_j \prod_j)$ which appear in (7). In the above, ρ_j, $\prod_j$, y_j for $j > t$ are, of course, forecasts made at t of the respective real interest rates, relative user costs of durables and real non-property incomes prevailing in the future. In empirical work, as well as choosing tractable restrictions on preferences and so for (8), some assumptions need to be made about how consumers make these forecasts.

Considering the forecasts necessary to construct the relative user cost of durables $\prod_t$ immediately raises a potential problem. In recent experience, the variability in *ex post* $\prod_t$ has been tremendous. In the 1970s in most Western economies, the *ex post* real rate of interest r_{t+1} was frequently negative and, particularly when accompanied by relative price declines in durable goods, it is quite likely that *ex post* the user cost $\prod_t$ was sometimes negative. It is true that there are expectations mechanisms such as the adaptive one which, combined with a low adjustment parameter, might have made the *ex ante* user cost less variable than the *ex post* and perhaps prevented it from becoming negative. However, in practice, even fairly crude extrapolations of past experience would have led to a highly variable series for the relative user cost $\prod_t$, much more so than is plausibly consistent with the relative smooth behaviour of purchases of durables if (8) were the true model. In Muellbauer (1981), I tested a model based on (8) for British quarterly data. The evidence strongly rejects such a model both because of the failure of crossequation restrictions and because of the obvious statistical misspecification of the durables equation.

It seems that one needs a theory which gives rise to much more sluggishness or 'persistence' in purchases of durables. The standard way of building in such persistence is to posit adjustment costs for durables. There has been much work over the years (see Stone and Rowe 1957; Chow 1957; and Nerlove 1957), on the stock adjustment model which is an *ad hoc* way of building in sluggish adjustment. Weissenberger (1984) has estimated a model for quarterly British data on non-durable

and durable demands based on a quadratic utility function with quadratic adjustment costs and rational expectations. Apart from the coefficients on financial assets he finds a reasonable degree of coherence between the two equations and the durables equation fits well and its residuals are well behaved. Muellbauer and Pashardes (1982, revised 1987) build persistence effects into zpreferences in a rather more general way though, in the context of a quadratic utility function, the effect is similar to Weissen-berger's. Since their empirical results on annual British data for a system of nine demand functions, one of which is for durables, are rather satisfactory, it is worth examining the approach.

They assume that the utility function is intertemporally separable in *transformed* quantities z_{ij}:

$$u_t = V\left[v_t(z_{1t},\ldots,z_{nt}),\ldots,v_T(z_{1T},\ldots,z_{nT}),A_T\right]. \quad (9)$$

where

$$\begin{aligned} z_{ij} &= \left(S_{ij} - a_i S_{ij-1}\right)(\delta_i/1-a_i) \\ &= \left[q_{ij} + (1-\delta_i-a_i)S_{ij-1}\right](\delta_i/1-a_i). \end{aligned} \quad (10)$$

and where the stock $S_{ij} = q_{ij} + (1-\delta_i)q_{ij-1} + (1-\delta_t)^2 q_{ij-2} + \cdots$. In a steady state $S_i = q_i/\delta_i$ and so $z_i = q_i$ In the case of a non-durable good (for which $\delta_i = 1$),

$$z_{ij} = \frac{q_{ij} - a_i q_{ij-1}}{1-a_i}.$$

The parameter a_i can be thought of as a 'habit' or 'persistence' parameter when $0 < a_i < 1$. With diminishing marginal utility, the more the consumer consumed of good i last period the greater is his or her marginality utility and so 'need' for the good this period. When $a_i < 0$, it can be interpreted as a shortlived durability parameter since, with $\delta_i = 1$, a purchase last period but not earlier gives utility this period.

It was Spinnewyn (1979, 1981) who first extended the user cost concept of price of a durable good to more general transformed quantities of the type defined in (10) above. As long as these are linear functions of $(q_{ij}, q_{ij-1}, q_{ij-2}, \ldots)$ a user cost price p_{ij}^{υ} can be defined. Then, since the intertemporally separable form of the utility function permits two-stage budgeting (see Gorman, on SEPARABILITY below), there exist demand functions

$$z_{it} = g_i\left(p_t^{\upsilon}, x_t^{\upsilon}\right) \quad (11)$$

Where

$$x_t^{\upsilon} \equiv \sum_{i=1}^{a} p_{it}^{\upsilon} z_{it}.$$

From (10), purchases of good i

$$q_{it} = [a_i - (1-\delta_i)]S_{it-1} + \frac{1-a_i}{\delta_i} z_{it}. \quad (12)$$

This can also be written in the form

$$\begin{aligned} \Delta q_{it} = \frac{1-a_i}{\delta_i}\Delta z_{it} \\ + (1-a_i)(z_{it-1} - q_{it-1}). \end{aligned} \quad (13)$$

This elegantly expresses an extended kind of partial adjustment model termed an 'error correction model' by Hendry. Changes in q respond to changes in z but with a stabilizing feedback to last period's deviation between q and z. The lower is durability, i.e. the higher is δ_i and the higher is a_i, slower is the speed of adjustment. This makes it clear that in the absence of persistence effects, one should expect greater volatility of purchases for durables than for non-durables. In the empirical results reported for annual British data by Muellbauer and Pashardes, there are strong persistence effects both for durables and for non-durables though generally a little larger for durables. Thus the volatility of purchases for durables is higher than for non-durables but much lower than it would be for a zero persistence parameter a_i.

Equation (13) can be estimated in two ways. The one adopted by Muellbauer and Pashardes uses the identity

$$x_t^v \equiv \sum_{i=1}^{n} p_{it} z_{it}.$$

Potentially it suffers from the possible correlation between a disturbance term added to (13) and the z_{it}'s embodied in the budget x_t^v The alternative is to solve the intertemporal optimization problem to give a solution for x_t^v as a function of life-cycle wealth and price and rate of return expectations. Under this alternative some assumptions, as in the study of the life-cycle consumption function, need to be made to model expectations empirically and the results are likely to be sensitive to which assumptions are made.

Models of this type give good results for aggregate time series data, at least relative to formulations that ignore persistence or durability. But one may well wonder about the source of persistence to which the above analysis gives simple expression. In many ways the most plausible explanation is the gap between buying and selling prices of durables due to installation costs, transactions costs or to information asymmetries. According to Akerlof's (1970) 'lemon effect' the potential buyer of a used durable fears that the reason the owner wishes to sell is that the durable is a lemon. Since there may be no way that the seller can convince the buyer that it is not, a car that is only a few weeks old and has suffered no physical deterioration is likely to be saleable only at a price substantially below the showroom price. Thus even if a consumer expects substantial capital gains on a durable, there is little incentive to buy in order to sell at a profit. The most that can then be expected is some advancement of purchases that are likely to have been made soon anyway.

At the individual level, such a gap between buying and selling prices makes corner solutions likely (see Deaton and Muellbauer (1980), pp. 360–64 for a discussion). It is then very difficult to derive tractable econometric models that reflect the theory at all precisely. One might take something like (13) as a starting point and try to build in additional elements such as the asymmetric response that larger restrictions on selling than on buying suggest. For specific durables such as automobiles where most buyers of new cars trade in a used one, the differential of new and used car prices could be added to (13) as an extra, somewhat *ad hoc* explanatory variable.

There are alternative ways of modelling the effects of prices of complementary goods such as the gasoline needed to run a car. One way is just to include this price in the vector of prices p_t^v as long as the form of preferences embodied in $g_i()$ is quite general, e.g. a flexible functional form (Diewert 1971). But imposing rather more structure often yields rewards. If it is car services that appear in the utility function and the cost of these consists both of a user cost or rental equivalent and a running cost that depends on the fuel efficiency of the car, price effects are likely to be modelled more accurately. This kind of approach also suggests vintage effects that could be observable even on aggregate data. In response to the fuel price increases in the 1970s manufacturers eventually brought models to the market which were much more fuel efficient than previous vintages. The effective price differential between new models and the existing stock is likely to be an important element in the replacement decision. As the fuel efficiency gap between the existing stock and new models narrows, *ceteris paribus*, one would expect replacement demand to fall off. Similar vintage effects can arise through other quality improvements in new durables. The only way quality improvements could show up in the simple model leading to (13) is through falls in the quality adjusted price p^D.

Among the advantages of the assumption that new and used durables, when converted into efficiency units, are perfect substitutes is that it simplifies the analysis of the interaction of supply and demand. Total market supply of stock is the sum of the surviving stock $(1 - \delta_i)S_{it-1}$, which is given, plus new supply, which if firms are competitive is a function of current and expected values of p_i^D and costs of production. Aggregate demand for stock given (11) can be written in the form

$$S_{it} = a_i S_{it-1} + \frac{1 - a_i}{\delta_i} g_i\left(p_t^v, x_t^v\right).$$

If prices clear the market, equating supply and demand determines p_{it}^D which is thus endogenous

and ought to be treated as such in econometric work. In practice, however, new and used durables may not be perfect substitutes, new prices may be set by oligopolistic producers and the second-hand market, despite its imperfections, may be more demand responsive in its prices. Nevertheless, the above model is a bench-mark that raises issues which applied economists working on the demand for durables need to face.

The other source of data on the demand for the durables is household surveys. In such cross-sections, the discreteness of ownership must be explicitly recognized. The classic paper by Farrell (1954) was one of the first to analyse the threshold effects which govern ownership. The modern economic treatment in its simplest form can be explained as follows. Let P^D be the rental price of a durable good and the budget constraint be

$$pq + p^D S = x \tag{14}$$

Suppose $S = 1$ if the durable is owned and $S = 0$ if not. Let the single period utility function be

$$u = v(q, S, b, \in)$$

where b is a vector of observable household characteristics and ε summarizes unobservable ones in a scalar. If $S = 0$, we can solve for $q = x/p$ from the budget constraint and if $S = 1$, $q = (x - P^D)/P$. Thus the durable is owned if the utility from owning $v[(x - p^D)/p,\ 1,\ b,\ \in]$ exceeds that from not owning $v[x/p,\ 0,\ b,\ \in]$. These solved out utility functions are termed 'indirect utility functions'. This ownership criterion is still valid if, in fact, durables of this type vary in size, performance, luxuriousness or other characteristics that can be summarized in a quality index. $S = 1$ then refers to the minimum quality available while $S > 1$ for higher qualities. Maximizing (15) subject to (14) then gives conventional demand functions, given $S \geqq 1$, $q = g(x, p, p^D; b,\ \in)$, $S = g^D(x, p, p^D; b,\ \in)$ and the durable is owned if

$$g^D\left(x, p, p^D; b, \in\right) \geq 1. \tag{16}$$

Here p^D is a quality corrected price index, which in a single cross-section, like p, is usually assumed to be the same for all consumers. Equation (16) suggests the use of Probit or Logit analysis to examine ownership variations on micro-data (see McFadden 1973, McFadden 1981). Given information on rental expenditure defined as $p^D S = p^D g^D(x, p, p^D; b,\ \in)$ if $g^D(\ ) \geq 1$, Tobit analysis is the appropriate technique (see Tobin 1958).

Equation (16) also has implications for 'quasi'-Engel curves which link ownership in an income bracket to the income level. Suppose that (16) holds if $\in < \theta(x, p, p^D, b)$. Then the proportion of households with budget x and characteristics b owning the durable is

$$\int_{-\infty}^{\theta} f(x, b, \in)\mathrm{d}\in \Big/ \int_{-\infty}^{\infty} f(x, b, \in)\mathrm{d}\in \tag{17}$$

where $f(\ )$ is the joint probability density of x, b and ε. Provided the budget x can be plausibly linked to observable income, this provides a justification for the 'quasi'-Engel curves estimated, for example, by Aitchison and Brown (1957), Cramer (1962), Pyatt (1964), and Bonüs (1973). These studies have not always, however, given as much attention to the household characteristics b as they might have done. If b and ε were independently distributed of x and θ monotonic in x, then there must be a sigmoid relationship between the level of x and the proportion owning the durable given the sigmoid shape of the cumulative distribution function for x. In practice, though there is quite a high correlation between such household characteristics as size and income, empirical 'quasi'-Engel curves are usually sigmoid in shape. In aggregate, as average income rises over time there is also a sigmoid relationship between the average income level and the aggregate proportion owning a durable. As Deaton and Muellbauer (1980, pp. 370–1) note, this sigmoid shape could be partly due to the sigmoid shape of the cumulative distribution function of income and partly due to other causes of diffusion such as epidemic models of the spread of a disease which may be appropriate when new goods such as television are introduced.

For simplicity the discussion above has taken the case of a single type of durable. But as

McFadden (1981) and Dubin and McFadden (1984) demonstrate, the generalization to a portfolio of different kinds of durables still results in tractable models which can be estimated by maximum likelihood techniques. All these models of ownership, however, need to be given a long-run interpretation which abstracts from transactions costs and imperfections in second-hand markets. The latter may cause specific households to have ownership patterns that differ from those they would have in a steady state. In cross-sections, information on past decisions and past income is usually missing, so that the kind of dynamic elements discussed in the earlier part of this entry cannot be analysed empirically.

See Also

- ▶ Household budgets
- ▶ Housing Markets
- ▶ Separability

Bibliography

Aitchison, J., and J.A.C. Brown. 1957. *The lognormal distribution*. Cambridge: Cambridge University Press.

Akerlof, G. 1970. The market for 'lemons'. *Quarterly Journal of Economics* 84(3): 488–500.

Bonüs, H. 1973. Quasi-Engel curves, diffusion, and the ownership of major consumer durables. *Journal of Political Economy* 81(3): 655–677.

Chow, G. 1957. *Demand for automobiles in the US: a study in consumer durables*. Amsterdam: North-Holland.

Cramer, J.S. 1957. A dynamic approach to the theory of consumer demand. *Review of Economic Studies* 24: 73–86.

Cramer, J.S. 1962. *A statistical model of the ownership of major consumer durables*. Cambridge: Cambridge University Press.

Deaton, A., and J. Muellbauer. 1980. *Economics and consumer behaviour*. Cambridge: Cambridge University Press.

Diewert, W.E. 1971. An application of the Shephard duality theorem: A generalized Leontief production function. *Journal of Political Economy* 79(3): 481–507.

Dubin, J.A., and D. McFadden. 1984. An econometric analysis of residential electric appliance holdings and consumption. *Econometrica* 52(2): 345–362.

Farrell, M.J. 1954. The demand for motor cars in the United States. *Journal of the Royal Statistical Society, Series A* 117(2): 171–201.

Fisher, I. 1930. *The theory of Interest*. New York: The Macmillan Company.

Jorgenson, D.W. 1963. Capital theory and investment behaviour. *American Economic Review, Papers and Proceedings* 53: 247–259.

McFadden, D. 1974. Conditional logit analysis of qualitative choice behavior. In *Frontiers in Econometrics*, ed. P. Zarembka. New York: Academic Press.

McFadden, D. 1981. Econometric models of probabilistic choice. In *Structural analysis of discrete data*, ed. C. Manski, and D. McFadden. Cambridge, MA: MIT Press.

Muellbauer, J. 1981. Testing neoclassical models of the demand for durables. In *Essays in the theory and measurement of consumer behaviour*, ed. A.S. Deaton. Cambridge: Cambridge University Press.

Muellbauer, J. and P. Pashardes 1982, *Tests of dynamic specification and homogeneity in a demand system*. (Discussion paper 125, Birkbeck College, 1982; revised as Institute of Fiscal Studies, discussion paper, 1987).

Nerlove, M. 1957. A note on long-run automobile demand. *Journal of Marketing* 21(July): 57–64.

Pyatt, G. 1964. *Priority patterns and the demand for household durable goods*. Cambridge: Cambridge University Press.

Spinnewyn, F. 1979. The cost of consumption and wealth in models with habit formation. *Economics Letters* 2(2): 145–148.

Spinnewyn, F. 1981. Rational habit formation. *European Economic Review* 15(1): 91–109.

Stone, R., and D.A. Rowe. 1957. The market demand for durable goods. *Econometrica* 25(July): 423–443.

Tobin, J. 1958. Estimation of relationships for limited dependent variables. *Econometrica* 26(January): 24–36.

Weissenberger, E. 1984. *An intertemporal system of dynamic consumer demand functions*. Centre for Labour Economics, London School of Economics, Discussion paper No. 186.

Consumer Expenditure

Angus Deaton

Abstract

Consumers' expenditure is a central concern of economics, both in microeconomic terms (the relationship between prices, expenditure and welfare) and in macroeconomic terms (the relationship between expenditure and income).

This article examines the interplay between theory and evidence in the study of consumers' expenditure and its composition. Although models have been developed from the theory of consumption that illuminate much of the available data, many standard presumptions of economics lack substantial bodies of evidence such as central theories in the natural sciences enjoy.

Keywords

Almost ideal demand system (AIDS); ARIMA process; Bernoulli utility functions; Compensated demand; Consumer expenditure; Consumers' expenditure; Consumption function; Convexity; Demand curve; Demand functions; Denver Income Maintenance Experiment; Econometrics; Elasticity; Engel curves; Engel's law; Euler equation; Fisher, I.; Fixed needs model; Flexible functional forms; Friedman, M.; Generalized axiom of revealed preference (GARP); Generalized Leontief system; Hicks, J.; Household budgets; Inflation; Intertemporal utility functions; Keynes, J.; King, G.; Kuznets, S.; Law of demand; Life cycle hypothesis; Life-cycle income model; Linear expenditure system; Lucas, R.; Marginal rate of substitution; Michigan Panel Study of Income Dynamics (PSID); Modigliani, F.; Non-linear optimization; Optimal taxation; Permanent income model; Prices; Ramsey, F.; Random walk model; Revealed preference theory; Risk aversion; Roy's identity; Samuelson, P.; Saving ratio; Shephard's lemma; Slutsky matrix; Stigler, G.; Stone, J.; Stone-Geary utility function; Substitution effect; Tax reform analysis; Unit root model; Utility maximization

JEL Classifications

E2

The study of consumers' expenditure, both in total and in composition, has always been of major concern to economists. Neoclassical economics sees the delivery of individual consumption as the main object of the economic system, so that the efficiency with which the economy achieves this goal is the criterion by which alternative systems, institutions and policies are to be judged. Within a capitalist economy, such considerations lead to an examination of the relationship between *prices* and consumption behaviour, and theoretical development and empirical analysis have been a major continuous activity since the middle of the last century. Even older is the tradition of using individual household budgets to dramatize poverty, and the relationship between household incomes and household expenditure patterns has occupied social reformers, statisticians and econometricians since at least the 18th century. In more modern times, it has been recognized that the study of public finance and of taxation depends on a knowledge of how price changes affect the welfare and behaviour of individuals, and the recent development of optimal tax theory and of tax reform analysis has placed additional demands on our understanding of the links between prices, expenditures and welfare.

In the last fifty years, aggregate consumption has become as much of an object of attention as has its composition, and in spite of a common theoretical structure, there has been a considerable division of labour between macro economists, interested in aggregate consumption and saving, and micro economists whose main concern has been with composition, and with the study of the effects of relative prices on demand. The interest of macroeconomics reflects both long-term and short-term interests. What is not consumed is saved, saving is thrift and the basis for capital formation, so that the determinants of saving are the determinants of future growth and prosperity. More immediately, aggregate consumption accounts for a large share of national income, typically more than three-quarters, so that fluctuations in behaviour or 'consumption shocks' have important consequences for output, employment, and the business cycle. Since Keynes's *General Theory,* the consumption function, the relationship between consumption and income, has played a central role in the study of the macroeconomy. Since the 1930s, there has been a continuous flow of theoretical and empirical developments in consumption function research,

and some of the outstanding scientific achievements in economics have been in this field.

In this essay, the major themes will be the interplay between theory and evidence in the study of consumers' expenditure and its composition. If economists have any serious claim to being scientists, it should be clearly visible here. The best minds in the profession have worked on the theory of consumption and on its empirical implementation, and there have always been more data available than could possibly be examined. I hope to show that there have been some stunning successes, where elegant models have yielded far from obvious predictions that have been well vindicated by the evidence. But there is much that remains to be done, and much that needs to be put right. Many of the standard presumptions of economics remain just that, assumptions unsupported by evidence, and while modern price theory is logically consistent and theoretically well developed, it is far from having that solid body of empirical support and proven usefulness that characterizes similar central theories in the natural sciences.

A Simple Theoretical Framework

Almost all discussions of consumer behaviour begin with a theory of *individual* behaviour. I follow neoclassical tradition by supposing that such behaviour can be described by the maximization of a utility function subject to suitable constraints. The axioms that justify utility maximization are mild, see any microeconomic text such as Varian (1978/1984) or Deaton and Muellbauer (1980b), so that utility maximization should be seen as no more than a convenient framework that rules out the grossest kind of behavioural inconsistencies. The assumptions that have real force are those that detail the constraints facing individuals or else put specific structure on utility functions. Perhaps the most general specification of preferences that could be considered is one that is written

$$u_t = E_t\{f(q_1, q_2, \ldots, q_t, \ldots, q_T)\} \qquad (1)$$

where u_t is utility at time t, E_t is the expectation operator for expectations formed at time t, q_1 to q_T are vectors of consumption in periods 1 to T, and $f(\cdot)$ is a quasi-concave function that is non-decreasing in each of its arguments. Several things about this formulation are worth brief discussion. The function $f(\cdot)$ yields the utility that would be obtained from the consumption vector under certainty, and it represents the utility from a *life-time* of consumption; the indices 1 to T therefore represent *age* with 1 the date of birth and T that of death. The expectation operator is required because choice is made subject to uncertainty, not about the choices themselves, which are under the consumer's control, but about the consequences of current choices for future opportunities. It is not possible to travel backward through time, so that choices once made cannot be undone, and yet the cost of current consumption in terms of future consumption foregone is uncertain, as is the amount of resources that may become available at future dates. The consumer must therefore travel through life, filling in the slots in (1) from left to right as best as he or she can, and at time (or age) t, everything to the left will be fixed and unchangeable, whether now seen to be optimal or not, while everything ahead of t is subject to the random buffeting of unexpected changes in interest rates, prices, and incomes. The solution to this sort of maximization problem has been elegantly characterized by Epstein (1975); here I shall work with something that is more restrictive but more useful and note in section "Recent Econometric Experience" below some phenomena that are better handled by the more general model.

Intertemporal utility functions are frequently assumed to be *intertemporally additive*, so that the preference rankings between consumption bundles in any two periods or ages are taken to be independent of consumption levels in any third period. If so, the utility function (1) takes the more mathematically convenient form

$$u_t = E_t \sum_{r=1}^{T} v_r(q_r). \qquad (2)$$

Note that by writing utility in the form (2), since the expectation operator is additive over states of the world preferences are in effect assumed to be

simultaneously additive over both states and periods, an assumption that can be formally defended, see Gorman (1982) and Browning et al. (1985). It has the consequence that risk aversion and intertemporal substitutability become two aspects of the same phenomenon. Individuals that dislike risk, and will pay to avoid it, will also attempt to smooth their consumption over time and will require large incentives to alter their preferred consumption and saving profiles. Note also that the additive structure of (2) means that, unlike the case of (1), previous decisions are irrelevant for current ones. For decision-making at time t, bygones are bygones, and conditional on asset and income positions, future choices are unaffected by what has happened in the past. There can therefore be no attempt to make up for lost opportunities, nor can such phenomena as habit formation be easily modelled.

Because utility in (2) is intertemporally separable, maximization of life-time utility implies that, within each period, the period subutility function $v_t(\cdot)$ must be maximized subject to whatever total it is optimal to spend in that period. The period by period allocation of consumption expenditure to individual commodities need not, therefore, be planned in advance, but can be left to be determined when that period or age is reached, and period t allocation will follow according to the rule

$$\text{maximize } v_t(q_t) \text{ subject to } p_t \cdot q_t = x_t, \qquad (3)$$

where p_t is the price vector corresponding to q_t and x_t is the total amount to be spent in t. Problem (3) is one of standard (static) utility maximization, though note that x_t is not given to the consumer, but is determined by the wider intertemporal choice problem. Nevertheless, not the least advantage of the intertemporally additive formulation is its implication that the composition of expenditure follows the standard utility maximization rule. It allows separate attention to be given to demand analysis on the one hand, i.e. to the problem (3), and to the consumption function on the other hand, this being understood to be the intertemporal allocation of resources, i.e. the determination of x_t.

Write the maximized value of utility from the period t problem as $\psi_t(x_t, p_t)$, where $\psi(\cdot)$ is a standard indirect utility function. The original intertemporal utility function then takes the form

$$u_t = E_t \sum_{r=0}^{T-t} \psi_r(x_{t+r}, p_{t+r}). \qquad (4)$$

The constraints under which this function is maximized are most conveniently analysed through the conditions governing the evolution of wealth from period to period. If A_t is the (ex-dividend) value of assets at the start of period t, N_{it} is the nominal holdings of asset i with price P_{it}, d_{it} is the dividend on i paid immediately before the beginning of t, and y_t is income in period t, then

$$A_{t+1} = \sum_i N_{it}(P_{it+1} + d_{it+1}) \qquad (5)$$

$$\sum_i N_{it} P_{it} = A_t + y_t - x_t. \qquad (6)$$

Conditions (5) and (6) determine how wealth evolves from period to period, and the picture is completed by requiring that the consumer's terminal assets be positive, i.e.

$$A_{T+1} \geq 0 \qquad (7)$$

To solve this problem, the technique of backward recursion is used. This rests on the observation that it is impossible to know what to do in period t without taking into account the problem in period $(t + 1)$, nor that in $(t + 1)$ without thinking about $(t + 2)$, and so on. However, in period T there is no future, so that looking ahead from date t, we can write subutility in period T in terms of that period's price and inherited assets, and we write this as v_T, i.e.

$$v_T = v_T(A_T) = \psi_T(A_T + y_T, p_t). \qquad (8)$$

Given this, the consumer can look ahead from period t to period $(T - 1)$ and foresee that the problem then will be to choose the composition of assets N so as to maximize v_{T-1}, where

$$v_{t-1}(A_{T-1}) = \max_{N} \left[\!\left[\psi_{T-1}(A_{T-1} + y_{T-1} - N \cdot P_{T-1},\ p_{T-1}) + E_{T-1}\{v_T[N \cdot (P_T + d_T)]\} \right]\!\right]. \tag{9}$$

At the next stage, assets in $(T-2)$ will be allocated so as to trade off the benefits of consumption in $(T-2)$ versus the benefits of A_{T-1} in v_{T-1} in (9) above and again yielding a maximized value v_{T-2}. As we follow this back through time, the consumer finally reaches the current period t, where he or she faces an only slightly complicated version of the usual 'today tomorrow' trade-off; the asset vector N must be chosen to solve the problem,

$$u_t = \max_{N} \psi_t(A_t + y_t - N \cdot P_t, p_t) + E_t\{v_{t+1}[N \cdot (P_{t+1} + d_{t+1})]\}. \tag{10}$$

From this sequence of problems, several important results readily follow. First, consider the derivatives of each of the functions $v_r(A_r)$ which represent the marginal value of an extra unit of currency for the remaining segment of life time utility from r through to T. By the envelope theorem (see for example Dixit (1976) for a good exposition), it is legitimate to differentiate through the maximization problem, from which

$$v_r'(A_r) = \partial\psi_r/\partial x_r = \lambda_r, \text{say}, \tag{11}$$

so that λ_r is the marginal utility of money in period r. Secondly, the maximization of (10) with respect to portfolio choice gives the relationship, for each asset i,

$$P_{it}\partial\psi_t/\partial x_t = E_t\{(P_{it+1} + d_{it+1})\partial\psi_{t+1}/\partial x_{t+1}\} \tag{12}$$

which, defining the asset *return* R_{it+1} as $(P_{it+1} + d_{it+1})/P_{it}$, and using (11) can be rewritten in the simple form

$$\lambda_t = E_t(\lambda_{t+1}R_{it+1}). \tag{13}$$

This equation, in current parlance often referred to as the 'Euler equation', can be used to derive many of the implications of the theory of consumption. Note first that it is little more than the standard result that the marginal rate of substitution between today's and tomorrow's consumption should be equal to the relative price. However, the equation is set in a multiperiod framework, not a two-period one, and it explicitly recognizes the uncertainty in both asset returns and in the value of money in subsequent periods. The equation also holds for all i, i.e. for all assets, so that the result also has implications for asset pricing as well as for consumption and saving, and for this reason the model is often referred to as the consumption-asset pricing model. I shall return to these implications below.

The theory as presented above is the modern equivalent of the life-cycle theory of consumption that dates back to Irving Fisher (1930) and Frank Ramsey (1928), and that had its modern genesis in the papers by Modigliani and Brumberg (1954) and (1954, published 1979). Modigliani and Brumberg's treatment differs from the above only in not explicitly modelling uncertainty, and by including only a single asset. The modern version appears first in Breeden (1979) and in Hall (1978), see also Grossman and Shiller (1981).

Predictions and Evidence

One of the most important implications of the theory above, and of Eq. (13) in particular, is that the evolution of consumption over the life-cycle is independent of the pattern of income over the life-cycle. The asset evolution Eqs. (5) and (6) allow consumers to borrow and lend at will, so that the only ultimate constraint on their consumption is one of life-time solvency. In consequence, consumption patterns are free to follow tastes, the evolution of family structure, or the different needs that come with ageing, provided that in the end total life-time expenditure lies within (total) life-time resources, whether from inherited wealth or from labour income. It is often assumed that tastes are such that consumers prefer to have a relatively smooth consumption stream, and this can be illustrated from a special case of Eq. (13). Assume that the within-period utility function is

homothetic so that $\psi(x, p)$ is $\varphi(x/a(p))$ for some linearly homogeneous function $a(\cdot)$, and that $\varphi(\cdot)$ has the isoelastic from with elasticity $(1 - \sigma)$. Life-time utility takes the form

$$u_t = \sum_{r=0}^{T-t} (1+\delta)^{-r} \left[x_{t+r}/a\left(p_{t+r}\right)\right]^{1-\sigma} \qquad (14)$$

where δ is the rate of pure time preference, and $\sigma \geq 0$ is the coefficient of relative risk aversion and the reciprocal of the intertemporal elasticity of substitution. Equation (14) can be used to evaluate (13), and gives immediately

$$E[\{(1 + r_{t+1})/(1+\delta)\}\{c_t/c_{t+1}\}^{\sigma}] = 1 \qquad (15)$$

where r_{t+1} is the real after tax rate of interest from t to $t + 1$ on any asset, and c_t is real consumption, $x_t/a\,(p_t)$. Equation (15) shows that, if expectations are fulfilled, consumption will grow over the life-cycle if the real rate of interest is greater than the rate of pure time preference, and vice versa, while with $r_t = \delta$, consumption is constant with age. These results are of course an artefact of the specific assumptions about utility, and for any real household consumption can be expected to vary predictably with age according to patterns of family formation, growth, and ageing; Modigliani and Ando (1957) have suggested that consumption per 'equivalent adult' might be constant over the life-cycle. But whatever the shape of preferences, there need be no relationship between the profiles of consumption and of income; income can be saved until it is needed, or borrowed against if it is not yet available.

Independent of the life-time *pattern* of consumption is its level, which under the life-cycle model is determined by the level of total life-time resources, so that individuals with the same tastes but with higher incomes or higher inherited assets will have higher levels of consumption throughout their lives. If the future were entirely predictable, the consumption plan at any point in time could be decided with reference to the level of total wealth, this being the value of financial assets and the discounted present value of current and future incomes. In this sense, the life-cycle model is a permanent income theory of consumption, where permanent income is the annuity value of lifetime wealth, though the lifetime interpretation is only one of the many that are offered in Friedman's (1957) original statement. Whether life-cycle or not, linking consumption to *future* incomes has important consequences. First, consumption will respond only to 'surprises' or 'shocks' in income; changes in income that have been foreseen are already discounted in previous behaviour and should not induce any changes in plans. Of course, this does *not* mean that consumption will not change along with changes in income; a change may have been planned in any case, and some proportion of any actual change may well have been unforeseen. However, if a substantial fraction of the regular changes in income over the business cycle are foreseen by consumers, or if unanticipated fluctuations in income are regarded as only temporary with limited consequences for total life time resources, then consumption will not respond very much to cyclical fluctuations in income. Aggregate consumption is indeed much smoother than is aggregate income, and this has been traditionally accepted as an important piece of confirmatory evidence. I shall take up the matter again below when I deal with the recent econometric evidence.

The distinction between measured income and permanent income is also important for the interpretation of cross-sectional evidence. Since measured income can be regarded as an error-ridden proxy for permanent income, the regression of consumption on measured income will be biased downward (rotated clockwise) compared with the true regression of consumption on permanent income. Crosssectional regressions, or time-series regressions of simple Keynesian consumption functions will therefore tend to understate the long-run marginal propensity to consume. Well before the work on life-cycle models, Kuznets (1946) showed that the long-run saving ratio in the United States had been roughly constant in spite of repeated cross-sectional analyses showing that the saving ratio rose with income, and the life-cycle theory could also readily account for these findings. It is interesting to note that the constancy of the saving ratio is far from being well

C

established as an empirical fact; the evidence for other countries with long-run data is very mixed, and even the United States saving ratio is clearly influenced in the long-run by technical change, migration patterns, and demographic shifts, see Kuznets (1962) and Deaton (1975). Life-cycle and permanent income theories also predict that households with atypically high income will tend to save a great deal of it, a prediction which explained the apparently anomalous finding that black households tend to save more than white households at the same level of measured income; since blacks typically have lower household income than whites, those with the same measured income can be expected to have a higher transitory component.

The Modigliani and Brumberg life-cycle story was also important because it offered a story of capital accumulation in society as a whole that relied on the way in which people made preparation for their own futures, particularly for their future retirement. In a stationary life-cycle economy, in which there is neither economic nor population growth, aggregate saving is zero, and the old, as they dissave, pass on the ownership of the capital stock to the next generation who are, in turn, saving for their own retirement. With either population or income growth, the aggregate scale of saving by the young would be greater than that of dissaving by the old, so that, to a first approximation, the aggregate saving ratio, while in the long run independent of the *level* of national income, would depend on the sum of its population and per capita real income growth rates. Modigliani (1986), in his Nobel address, has given an account of how very simple stylized models of saving and refinement yield quite accurate predictions of the saving ratio and of the ratio of wealth to national income, and the predictions about the growth effects have been repeatedly borne out in international comparisons of saving rates, see Modigliani (1970), Houthakker (1961, 1965), Leff (1969) and Surrey (1974). Perhaps the only problem with these interpretations is that there is little evidence that the old actually dissave, except by running down state social security or pension schemes; see for example Mirer (1979). Partly, this may be a rational response to uncertainty about the date of death and about possible medical expenses near the end of life (Davies 1980), partly there may be statistical problems of measurement (Shorrocks 1975), and partly consumers may wish to leave bequests. However, most countries' tax systems penalize donors who do not pass on assets prior to death, so the reason for the size of actual bequests remains something of a mystery. Bernheim et al. (1985) have gone so far as to suggest that parents retain their wealth until death in order to control their heirs and to solicit attention from them. They claim empirical support for a positive relationship between visits by children to their parents and parents' bequeathable assets; visits are apparently especially frequent to rich sick parents, but not at all frequent to poor sick parents. Related to the dispute about the reason for bequests is a parallel dispute on their importance in the transmission of the capital stock, see the original contribution by Kotlikoff and Summers (1981) and Modigliani's reply, summarized in his (1986) Nobel lecture.

The life-cycle and permanent income models also provided the econometric specifications for a generation of macroeconometric models. Ando and Modigliani (1963) suggested a simple form for the aggregate consumption function in which real aggregate consumption was a linear function of expected real labour income, *YL,* and of the real value of financial assets, i.e.

$$c_t = \alpha E_t(YL) + \delta W_t. \tag{16}$$

In practical econometric work, the expectation was typically replaced by a linear function of current and past values of labour income, a procedure that can be formally justified by modelling labour income as a linear ARIMA process, a topic to which I shall return below. Wealth or a subset of wealth was included as data allowed, although sometimes the return to wealth was included with labour income which could then be replaced by total income, so that, with smoothing, (16) becomes a permanent (total) income model of consumption. A favourite variant, suggested in Friedman (1957), was to model permanent income as an infinite moving average of current income with geometrically declining weights,

$$y_t^p = (1-\lambda)\sum_{r=0}^{\infty}\lambda^r y_{t-r}, \tag{17}$$

so that if current consumption is proportional to permanent income, substitution yields

$$c_t = kc_{t-1} + k(1-\lambda)y_t, \tag{18}$$

a formulation that is also easy to defend if consumers 'partially adjust' to changes in current income. Models like (18), possibly with additional lags, and with the occasional appearance of more or less 'exotic' regressors, such as wealth, interest rates, inflation rates, money supply, as well as various dummy variables for 'problem' observations, were the standard fare of macroeconometric models in their heyday, from the early sixties for about a decade and a half. They fit the data well, they accounted for the smoothness of consumption relative to income, and they accorded at least roughly with the general features of the life-cycle and permanent income formulations which provided them with pedigree and general theoretical legitimacy. Dozens of papers could be cited within this tradition; those by Stone (1964, 1966), Evans (1967), and Davidson et al. (1978) will perhaps stand as good examples.

Recent Econometric Experience

In the mid-1970s, the general state of complacency of macroeconomic modelling was rapidly eroded, largely by the apparent inability of the standard models to explain, let alone to predict, the coexistence of unemployment and inflation. The relationship between consumption and income did not escape some of the blame, although the main focus of attack was elsewhere. Standard consumption functions, which had worked well into the early seventies, seriously under-predicted aggregate saving during the period of (at least relatively) rapid inflation that characterized most Western economies in the middle of the decade. The implementation of the theory of the consumption function was also singled out for discussion in Lucas's famous (1976) essay that became known as the Lucas 'critique'. As Lucas forcefully argued, if consumption is determined by the discounted present value of *expected* future incomes, the response of consumption to a change in income is not well-defined until we know how expectations of income are formed. Each observed realization will cause a re-evaluation of future prospects in accordance with formulae that depend on the nature of the stochastic process governing income. If the nature of the stochastic process is changed, for example by a fundamental change in the tax code, then the way in which information is processed will change, and new information about incomes will have different implications for future expectations and for future consumption. This insight is of great importance, although its implications for econometric modelling were initially taken much too negatively; if the rules keep changing, econometric models will be inherently unstable (as evidenced by their performance in the mid-seventies) and we should give up trying to find stable relationships. Instead, as events have shown, the introduction of rational expectations has given a whole new lease of life to the study of consumption, with developments as positive as anything that has happened since the life-cycle and permanent income models were the 'new' theories in the mid-fifties. Lucas's critique suggested at least two lines for research. First, could the failure of consumption functions, or indeed of macroeconometric models in general, really be traced to a change in the way expectations were formed? If so, it ought to be possible to detect changes in the stochastic process generating real income. Second, and more generally, if expectations are important, there ought to be high returns to the simultaneous modelling of consumption and income, so that knowledge of the structure of the latter can be used either to estimate the consumption function or to test for the validity of the expectations mechanism. My own reading of the evidence is that the Lucas critique is *not* capable of explaining the failure of the empirical consumption function, but that the under-prediction of saving resulted from ignorance of the fact that saving appears to respond positively to inflation, or at least to unanticipated inflation.

C

There is overwhelming evidence from a large number of countries, see in particular Koskela and Viren (1982a, b), that saving increased with inflation in the 1970s, even when we allow for real income and its various lags. Such a finding is also consistent with the life-cycle theory since unanticipated inflation imparts a negative shock to real assets, so that risk-averse, low inter-temporal elasticity consumers will save to replace the lost assets so as to avoid the chance of low consumption later. It is also possible to explain the relationship through the confusion between relative and absolute price changes that is engendered by unanticipated inflation in an environment in which goods are bought sequentially, see Deaton (1977), but it would be hard to devise a test that would separate this from the life-cycle explanation. But if inflation was indeed the cause of the failure of the empirical consumption functions, then it is a standard enough story. An important variable was omitted from the analysis, it had not been very variable in the past so that its omission was hard to detect, and economists had not been imaginative enough to perceive its importance in advance. The Lucas critique is only one of the many problems that can beset an econometric equation, and it does not seem to have been the fatal one in this case.

The second research direction, the joint examination of income and consumption, has proved more productive. The first important step was taken by Hall (1978), who pointed out that Eq. (15) implies that, as an approximation consumption should follow a random walk with drift. To see why, assume that the real interest rate r is constant and known, and write (15) in the form

$$c_{t+1}^{-\sigma} = \{(1+\delta)/(1+r)\}c_t^{-\sigma} + \varepsilon_{t+1} \qquad (19)$$

where the expectation at t of ε_{t+1} is zero. Equation (19) is exact, but a convenient expression can be reached by factoring c_t out of the right hand side, taking logarithms, and approximating. This gives

$$\ln c_{t+1} = \ln c_t + g + v_{t+1} \qquad (20)$$

where g is positive or negative as r is greater than or less than δ, and the 'innovation' v_{t+1}, like ε_{t+}, has expectation zero at time t. Equation (20) shows that, in the absence of 'news', consumption will grow or decline at a steady rate g, so that nothing that is known by the consumer at time t or earlier should have any value for predicting the deviation of the rate of change of consumption from its constant mean. The result is often referred to as the 'random walk' property of consumption, though the theory does not predict that v_{t+1} has constant variance, so that, strictly speaking, the stochastic process is not a random walk.

For someone used to thinking about the consumption function as the relationship between consumption and income, Eq. (2) is notable for the apparent absence of any reference to income. But of course income can appear through the stochastic term v_{t+1} if current income contains new information about its own value or about future values of income, and this will generally be the case. The random walk model does not predict that consumption should not respond to current income. It does however predict that, conditional on lagged consumption, past income or changes in income should not be correlated with the current change in consumption, and a considerable amount of effort has recently gone into testing this proposition. In Hall's (1978) original paper, to the surprise of the author and of much of the profession, the model worked well for an aggregate of United States consumption of nondurables and services. The level of consumption certainly depends on its own lagged value, but the addition of one or more lagged values of income or of further lagged values of consumption did not significantly add to the explanatory power of the model. Hall examined the role of the number of other lagged variables and discovered that lagged stockmarket prices had predictive power for the change in consumption, so that he concluded by formally rejecting the model. However, the overwhelming impression was favourable, at least relative to expectations.

Hall's test procedures are attractive because they do not depend on the properties of the income process, and focus only on consumption and its lags. But robustness comes at the price of power, and later work has devoted considerable attention to the joint properties of consumption and real

income. Perhaps the natural route to modelling is to find a representation of real income as a stochastic process, typically as some sort of ARIMA. Once this is known, changes in income can be decomposed into anticipated and unanticipated components using the standard forecasting formulae from statistical time series analysis, so that it becomes possible to test whether consumption responds to one but not to the other. The random walk model seemed not to survive these tests so well. Papers by Flavin (1981) and by Hayashi (1982) showed that, for United States data, consumption is sensitive to *anticipated* changes in income, something that should not be the case in a thoroughgoing life-cycle model in which consumers are efficiently looking into the future. The phenomenon became known as the 'excess sensitivity' result, and was typically ascribed to the existence of a substantial number of consumers who wish to borrow against future income but are unable to do so. Such liquidity constrained consumers can be expected to consume all their available income, so that their consumption will increase one for one with all income changes, whether anticipated or not.

However, it is not clear that the excess sensitivity finding is itself robust. First, it is becoming increasingly recognized that the problems of econometric testing in the time-series models are more severe than had been generally supposed. The time series of both consumption and income are non-stationary, and it sometimes seems as if hypothesis testing in models involving non-stationary variables is like building on shifting sands; see Mankiw and Shapiro (1985, 1986) and Durlauf and Phillips (1986) for some of the problems. Second, there are a large number of variables other than income which can affect consumption, so that, according to (20), surprises in wealth and in inflation should affect consumption, as should the level of real interest rates. Adding even a few of these variables reduces degrees of freedom and diminishes the probability of being able to reject the basic model. Both Bean (1985) and Blinder and Deaton (1985) find that time-series models of consumption with several variables are more easily reconciled with the theory than are the simple two variable models. Not all of this should be ascribed to lack of degrees of freedom; for example Blinder and Deaton consistently find that unanticipated changes in wealth affect consumption and that anticipated changes do not. Third, even in a bivariate income-consumption model, Campbell (1987) has found that the model is largely consistent with the time-series evidence. Campbell recognizes the possibility of time-series feedback from lagged consumption to income, and models saving and the change in income as a bivariate vector-autoregressive system in which each series is regressed on lagged values of both. The structure of this representation then turns out to be very close to what it would have to be if the life-cycle rational expectations model were correct. The conflict between Campbell's results and the excess sensitivity findings are presumably accounted for by the feedback from saving to changes in labour income, since his model is otherwise compatible with the earlier ones.

Similarly mixed findings are also being uncovered from longitudinal panels that follow individual households over time. In contrast to the situation with labour supply, there are few panel data in the United States that cover household consumption, and most work has used the data on expenditure on food that is contained in the Michigan Panel Study of Income Dynamics (PSID). In an elegant paper, Hall and Mishkin (1982) found results that were in accord with the excess sensitivity results; there is a strong negative correlation in their data between changes in consumption and changes in lagged income that is inconsistent with the view that only surprises in income should matter. However, since in their data changes in income are negatively correlated over time, a negative correlation between the lagged income change and the change in consumption can be interpreted as a positive correlation between consumption changes and changes in actual income, as predicted by the model of liquidity constraints. Hall and Mishkin conclude that these results would be consistent with a model in which about one fifth of consumers were unable to borrow as much as they wished. Once again, these results were supported by other similar evidence, see in particular Zeldes (1985) and

Bernanke (1984), also using the PSID, Runkle (1983), using data from the Denver Income Maintenance Experiment, and Hayashi (1985a) using panel data from Japan. However, one potential problem with the use of panels is the importance of errors of measurement in such data. There is a considerable body of evidence that PSID income changes are subject to very substantial reporting errors, see in particular Altonji (1986), Duncan and Hill (1985), and Abowd and Card (1985). Altonji and Siow (1985) have recently estimated a model similar to Hall and Mishkin's using the PSID but with allowance for measurement error, and they find little conflict with the view that consumption responds only to news. However, it is unclear, at least to this reader, whether the acceptance of the model represents low power once errors of measurement are allowed for, or whether such errors really offer a plausible explanation for Hall and Mishkin's findings.

A more formal line of research has attempted to estimate the Euler condition (15) directly, thus avoiding the approximations made by Hall and by others. Rewrite (15) once more, this time as

$$(1 + r_{t+1})(c_{t+1})^{-\sigma} - (1 + \delta)(c_t)^{-\sigma} = \varepsilon_{t+1} \quad (21)$$

where, as before ε_{t+1} is orthogonal to any variable known in period t or earlier. Hansen and Singleton (1982) proposed that the parameters in (21) be estimated by a generalized methods of moments scheme. Suppose that we have two variables or instruments z_{1t} and z_{2t}, each known at time t, so that we have $E_t(z_{it}\ \varepsilon_{t+1}) = 0$ for $i = 1, 2$. We can then estimate the two unknown parameters, σ and Δ, by equating sample and theoretical moments, and solving the two equations, $i = 1, 2$

$$T^{-1}\sum_{t=0}^{T-1}[z_{it}\{(1 + r_{t+1})(c_{t+1})^{-\sigma} - (1 + \delta)(c_t)^{-\sigma}\}] = 0. \quad (22)$$

If, as is typically the case, we have more than two z-variables, then it will not generally be possible to choose the two parameters so that (22) is exactly zero. Instead, the vector can be made as small as possible, or more specifically, the parameters can be estimated by minimizing a quadratic form that can be thought of as a weighted sum of squares of the left-hand side of (22); see Hansen and Singleton for details. If the model were true, this minimized value ought to be small, so that with more instruments than parameters, the generalized method of moments procedure yields a test-statistic that is diagnostic for model adequacy.

Test procedures based directly on the Euler conditions have several notable advantages. As was the case for Hall's procedures, few assumptions have to be made about the structure of the income process, and the model satisfies the best professional standards of seeking a direct confrontation between theory and data with as few approximations and supplementary assumptions as possible. The model can also be readily extended to test the implications of the consumption asset pricing model by repeating the tests using the returns on a range of alternative assets, see (13) above. Hansen and Singleton's study, as well as several others, find that the test statistics are much too large to be consistent with the theory and so reject the intertemporal model implied by the Euler conditions. Given the apparent superiority of the tests, these results have been accorded a great deal of weight in the literature. However, while I believe that Hansen and Singleton's work represents a very important methodological advance, I think that there are good reasons for not treating their results as a definitive rejection of life-cycle theory. The high level of technique that is embodied in deriving the Euler equation, not to mention the complexity of generalized methods of moments estimation, should not blind us to the very simple, even simple-minded, economic story that underlies these models. Fundamentally, the Euler equation says that the marginal rate of substitution between today's and tomorrow's consumption should be equal to the rate of return on assets between today and tomorrow, so that estimation of the Euler equation, unlike the Hall or excess-sensitivity tests, focuses very directly on the relationship between real interest rates and changes in real consumption, and the model will not fit the data if there is no close association between the two. And it only takes a very cursory

inspection of United States time-series data to see that there is no such association. Real consumption grew in all but one year between 1954 and 1984, while real after-tax interest rates were as often negative as positive, so that consistency with the theory would require that the pure rate of time preference be negative. Nor is there any association between the rate of growth of consumption and the level of real after-tax interest rates, see Deaton (1986b) for some data. But this in no way reflects badly on the life-cycle theory. As was made perfectly clear in the original Modigliani and Brumberg papers, and it is the *essence* of the life-cycle model, aggregate consumption cannot be expected to behave like individual consumption. Imagine a stationary economy with neither population nor real income growth, in which there is an excess of real interest rates over the rate of pure time preference, and in which all consumers have identical additive lifetime preferences with isoelastic subutility functions. In such an economy, each individual has a consumption path that is growing over time, but aggregate consumption is constant, a result that is achieved by old people dying and being replaced by young people who have much lower consumption levels relative to their incomes. Unless we believe that there is some automatic and immediate relationship between real interest rates, time preference and growth, as would obtain for example along a 'golden age' growth path, or unless we believe that consumers have infinite lives, then there is no reason at all to suppose that aggregate consumption should look at all like the life-cycle path of a representative consumer. Representative agent models are frequently useful, and it is not very constructive to dismiss macroeconomics because it requires implausible aggregation assumptions. However, the life-cycle model provides a well-worked-out account of individual and aggregate saving, an account that is consistent with a good deal of other evidence and theory, and it *does not* predict that aggregate consumption should be consistent with the intertemporal optimization conditions for a single individual. The general question of the effects of interest rates on consumption is something that has remained in dispute for a long time, and in spite of repeated attempts to isolate the effect, careful studies have tended to be unable to do so, or at least to find effects that are at all robust, or that can be replicated on even slightly different data sets or data periods. Economic theories or policy prescriptions that rely on intertemporal substitution of consumption in response to changes in real interest rates are not well-buttressed by any solid body of empirical evidence.

Another useful approach to testing the life-cycle model is to consider the stylized facts of the income and consumption processes, and to see whether consumption behaves in the way that is to be expected given the stochastic process of income. Most people who have studied the time series for quarterly real disposable income in the United States agree that, like GDP, the series can be parsimoniously described by a model that is linear in its first two lags, i.e. an autoregression of the form

$$y_t = \alpha_1 + \alpha_2 y_{t-1} + \alpha_3 y_{t-2} + u_t \qquad (23)$$

where u_t is the income innovation, that part of current income that cannot be anticipated from previous observation of the series. Of course, real income is not a stationary series, but has a strong upward trend, and there is considerable disagreement about the nature of this trend, what is the economic story behind it, and how it should be modelled. One possibility is that real income contains a *deterministic* time trend, so that there is some sort of equilibrium growth path that cannot be altered by shocks to the economy. Shocks certainly exist, but they cause only short term temporary deviations from the path and have little or no long-term temporary deviations from the path and have little or no long-term significance. In this view, Eq. (23) applies to the *deviations* of income from trend, not to income itself; equivalently, (23) can be modified by including a linear or quadratic time trend. The alternative view is that there is no deterministic trend, but that the rate of change of income is a stationary stochastic series with constant mean. In practice, this can look very like the previous model, but there is the vital conceptual difference that in the second, non-deterministic model, there is nothing that will

ever bring income back to any deterministic path. In consequence, shocks to current income have permanent and long-lasting effects. The version of (23) that corresponds to this view can be written.

$$\{(y_t - y_{t-1}) - \gamma\} = \rho\{(y_{t-1} - y_{t-2}) - \gamma\} + u_t \tag{24}$$

which can readily be seen to be a special case of (23), though note that it is the case where the time series possesses a unit root, or is stationary in first differences. For (24) to be a valid specialization of (23), the quadratic equation with the α's of (23) as coefficients must have a unit root, hence the term. Equation (24) appears to fit the data well and the parameter ρ turns out to be around 0.4, so that (24) says that if the increase in real income in one quarter is greater than its long term mean, then the next quarter's increase is also likely to be above the mean, though by less. While the long-term mean of the rate of change of income is constant and equal to γ, good fortune (positive u's) and bad fortune (negative u's) never have to be paid for (or made up), since shocks are immediately consolidated into the income level, and growth goes on in the same way as before, but from the new base. As Campbell and Mankiw (1986) have emphasized, the unit root model exhibits shock *persistence,* while the deterministic trend model does not; they suggest that shock persistence is what we should expect if supply shocks predominate over demand shocks, with the reverse in standard Keynesian models where shocks are typically attributed to fluctuations in aggregate demand.

It turns out that it is almost impossible to tell these two processes apart on United States time-series data. Processes with unit roots are inherently difficult to tell apart from processes that are stationary around deterministic trends, and the tests that are available, Dickey and Fuller (1981), Phillips and Perron (1986), certainly cannot reject the hypothesis that (24) is a valid specialization of (23). Nor would the tests convince a believer in the deterministic model that income does not have a deterministic trend, even though it will readily be recognized that the deviations from trend are themselves close to non-stationarity. Since both process are special cases of (23) with the inclusion of a trend, and since each assumes parameter values that are very close to one another, one might think (and hope) that the two models would have very similar implications. But it is easy to see this is not true. If permanent income is taken as the annuity value of discounted future incomes, then (24) implies that any innovation u_t to current income, because it will persist forever, and because it can be expected to be followed by another infinitely persistent innovation of the same sign, will change permanent income by more than the amount of the innovation. Equation (25) below gives the formula for the change in permanent income, if the real interest rate is r, and if real income follows (24), see Flavin (1981) or Deaton (1986b),

$$\Delta y_t^p = \frac{(1+r)^2}{r+1-\rho} u_t \tag{25}$$

so that the change in permanent income is between one and a half and twice as large as the innovation in current income. By contrast, fitting the deterministic model yields a much smaller effect, with the change in permanent income about one fifth of the shock in measured income. Since consumption should change by about the same amount as does permanent income, the life-cycle model, together with the unit root formulation, yields the uncomfortable prediction that consumption should be *more* variable than income over the business-cycle, not less. If the unit root model is correct, then the life-cycle and permanent income models can be rejected because they predict what they were designed to predict, that consumption is smooth relative to real income! The deterministic model gives no such problems, but as yet we have no way of being sure that it is correct, unless, of course we assume from the start that the life-cycle story is true.

There is insufficient space in this essay to follow these issues further, or to discuss in detail the evidence for and against the two formulations of the stochastic process governing real income; the interested reader can refer to Deaton (1986b) and

to the evidence on persistence in GDP presented by Campbell and Mankiw (1986) and by Cochrane (1986). There are a number of possible solutions to these puzzles, and a great deal of empirical work remains to be done, though I suspect that the time-series data on income are insufficiently long to allow the isolation of the very long-run properties on which the permanent income theory rests, see in particular the interesting paper by Watson (1986).

Variations on the Basic Theme

There exist many interesting developments of the basic life-cycle model, and I have space to discuss only a few. I have already mentioned the role of liquidity constraints, and many people would take it as transparent that many consumers do not have access to unlimited credit, or else face borrowing rates that are higher than the rates at which they can lend. Of course, many consumers may be able to smooth their consumption without recourse to borrowing, and the borrowing needs of many others may be met by the typically rather good markets in home mortgages. For consumers who nevertheless wish to borrow but cannot, their spending will be closely tied to their actual income. For some of the theoretical and empirical literature on this point see Flemming (1973), Dolde and Tobin (1971), and Hayashi (1985b). The theoretical consequences of uncertainty about the date of death have been worked out by Yaari (1965), and as argued above, play a possibly important part in the explanation of the saving behaviour of the elderly.

Another line of research is the possible relaxation of the assumption that preferences are intertemporally additive. Allowing all periods (or ages) to interact with all other periods in an unrestricted way, as in Eq. (1), would be much too general to be useful, and the search has been for simple models that break the restriction in a natural and straightforward way. One useful analogy is with the theory of durable good purchases, where utility depends on the *stock* of assets possessed, the stock in turn being the integral of past purchases less depreciation. Purchases in one period therefore have consequences for utility in subsequent periods, something that will be taken into account by a forward looking consumer. In the case of durable goods, the assumption of perfect capital markets effectively converts durable into non-durable goods, with the price of a unit of stock for one period being the implicit rental or user cost, the latter being defined as the sum of interest cost, depreciation, and expected capital loss, see for example Diewert (1974) or Deaton and Muellbauer (1980b, ch. 13).

However, various authors, Houthakker and Taylor (1970) perhaps being the first, have extended the durable model to encompass 'psychic' stocks which, like physical stocks, are augmented by purchases and diminished by depreciation, but unlike physical stocks, can either increase or decrease utility. The latter case covers habit formation; consumption of an addictive good generates pleasure now, but engenders a hungry habit that is pleasureless but costly in the future. The model has been given an elegant formulation in two papers by Spinnewyn (1979a, b). As an example, see also Muellbauer (1985), take the utility function

$$u_t = \sum_{k=0}^{T-t} (1+\delta)^{-k} v(c_{t+k} - \alpha c_{t+k-1}) \qquad (26)$$

where α is a measure of habit formation. Spinnewyn maximizes this function with respect not to c_t, but with respect to the 'net' quantities $z_t = c_t - \alpha c_{t-1}$, and shows how to rewrite the budget constraint so as to define corresponding prices of the z's that reflect not only market prices of the goods, but also the costs of consumption now in terms of pleasure foregone later. Under certainty, and looking ahead from time t, the full shadow price of an additional unit of consumption now is

$$p_z = \sum_{k=0}^{T-t} [\alpha/(1+r)]^k p_{t+k} \qquad (27)$$

because the habits that are built up now have to be paid for later. Note that this sort of formulation also predicts that it is $c_t - \alpha c_{t-1}$ not c_t, that is

proportional to permanent income, so that consumption itself will adjust only sluggishly to changes in permanent income with habits causing a drag. Other formulations of non-separable preferences can be found in the papers by Kydland and Prescott (1982), and by Eichenbaum et al. (1984), both of which are concerned to reconcile fluctuations in the aggregate economy with the behaviour of a single representative agent.

Many of the models discussed so far assume that the consumption function actually exists, hence taking for granted the essentially keynesian assumption that income is given to the consumer, and is not chosen together with consumption. A considerable body of work has grown up in the last ten years that is concerned with the simultaneous choice of labour supply and consumption in a life-cycle setting. Heckman (1971) and Ghez and Becker (1975) are among the pioneers of this approach. Unlike the price of goods, the price of leisure tends to show a systematic pattern over the life-cycle, so that, if consumers are free to choose their hours, and if they can freely borrow and lend so as to transfer resources between periods, it will pay them to work hardest during those periods in their life-cycles when the rewards for doing so are highest, and to take their life-time leisure when wage rates are low and leisure is cheap. There is superficial evidence in favour of this story, and Ghez and Becker, followed by Smith (1977) and Browning et al. (1985), all find that workers tend to work longest hours in middle age when wage rates are high and the lowest number of hours at the beginning and end of the economically active life, when wage rates are relatively low. Consumption also tends to peak in the middle age, and this can be brought into the story by assuming that consumption and leisure are complements, so that the lack of leisure in middle age is partially compensated by high levels of expenditure. This elegant fable has also been made much of in equilibrium theories of the business cycle, which accounts 'unemployment' as a voluntary vacation taken when the real wage is low and leisure is on sale, see in particular Lucas and Rapping (1969) and Lucas (1981).

There now exists a growing volume of literature that shows just how much violence to the facts is done by this story. All the evidence quoted above looks across different individuals at different points in their life-cycles, while the theory says that the same individual will change his or her hours of work along with changes in the real wage over the life-cycle. Time-series and panel data from the United States and time-series of cross-sections from the United Kingdom suggest that this is simply not the case, see for example Mankiw et al. (1985), Ashenfelter and Ham (1979), Ashenfelter (1984), and Browning et al. (1985). Even MaCurdy's (1981) more postive study provides only very weak evidence, see in particular Altonji (1986). The joint consumption and labour supply story fares even less well than the labour supply model alone, and there is clear evidence that the way in which consumption and hours fluctuate over the cycle (sometimes together and sometimes in opposite directions) is not consistent with the way in which they move together over the life-cycle. The attempt to provide a unified theory of business and life-cycles has been an interesting and important one, but it cannot be said to have been successful.

I have been somewhat cavalier in my treatment of aggregation issues, choosing to emphasize them when I believe them to be important, for example in the fitting of Euler conditions, and ignoring them when it has been convenient to do so. Attempts to do better than this have not been notably successful. Formal conditions that allow aggregation in consumption function models are typically too restrictive to be useful, so that, in theory, changes in the distribution of income should have detectable effects on aggregate consumption. However, attempts such as that by Blinder (1975) to link the distribution of income to consumption have not been notably successful, perhaps because the income distribution is not variable, or because it changes smoothly enough over time to preserve a stable relationship between average income and average consumption. There is also an issue of aggregation over goods in order to define real consumption at all, even at the level of the individual agent. In the derivation in section "A Simple Theoretical Framework" above, I made the convenient assumption that within-period preferences were homothetic, so that an

index number of real consumption could be formed. But homotheticity, although very convenient for studying the consumption function, is very inconvenient for studying the allocation of expenditure among goods since it implies that the within-period total expenditure elasticities of each good are all equal to unity. Fortunately, there are aggregation results of Gorman's (1959), see also Deaton and Muellbauer (1980b, ch. 5) for an exposition that allows us to have the best of both worlds, at least if we remain with intertemporally additive preferences. If the single-period indirect utility function $\psi(x, p)$ takes the form known as the 'generalized Gorman polar form'

$$\psi(x,p) = F[x/a(p)] + b(p) \tag{28}$$

where $a(p)$ and $b(p)$ are linearly homogeneous functions of prices and $F(\cdot)$ is monotone increasing, then the real expenditure index $x/a(p)$ can serve as an indicator of real consumption just as in the homothetic case. This happens because when the consumer chooses the allocation of life-time expenditure over periods so as to maximize the intertemporal sum of terms like (28), the $b(p)$ terms are irrelevant. However, the intra-period demand functions that correspond to (28) do not display unitary elasticities unless the $b(p)$ is identically equal to zero, and quite general functional forms are permitted. There is therefore no real conflict between the analysis of the consumption function on the one hand, and the analysis of demand on the other. It is to the latter that I now turn.

Theoretical and Empirical Demand Functions

Demand functions are the relationships between the purchase of individual goods, income or total expenditure, prices, and a variety of other factors depending on the context. Economists have attempted to make empirical links between demand and price since Gregory King's famous demand curve for wheat, see Davenant (1699), and since the middle of the 19th century, there has been a great development in the theory of consumer behaviour. Much practical work continues in the tradition of King, paying little attention to formal theory, concerning itself instead with finding empirical regularities. For a firm studying the demand for its product, or for anyone interested in establishing a single price elasticity, this probably remains the best approach; the major developments in econometric technique and empirical formulation have not been much concerned with, or relevant to, these very practical questions. The pragmatic approach (the term comes from Goldberger's famous but unpublished (1967) study), probably reached its peak with the publication of Richard Stone's great monograph, (Stone 1954a), and much is still to be learned by a careful study of Stone's procedures for measuring income and price elasticities. However, in this essay, I shall follow the literature, and follow its more methodological approach.

The theory outlined in section "A Simple Theoretical Framework" above suggests that the demand functions of an individual consumer can be derived by maximizing a utility function $v(q)$ subject to a budget constraint $p \cdot q = x$, where x is total expenditure. In the analysis here, x is chosen at some previous level of decision making, but traditionally it is treated as if it were a datum by the consumer, the utility maximization yields a vector q that is some function $g(x, p)$, say, of total expenditure and prices. These demand functions cannot simply be any functions, but must have certain properties as a result of their origins in utility maximization. Obviously, the total value of the demands should be equal to total outlay x, the 'adding-up' property, and it must be true that proportional changes in x and in p do not have any effect on quantities demanded, the 'homogeneity' or 'absence of money illusion' property. Somewhat less obvious are the famous symmetry and negativity properties. These apply to the Slutsky (1915) matrix, S, the typical element of which is defined as

$$s_{ij} = \partial q_i/\partial p_j + q_i \partial q_j/\partial x. \tag{29}$$

As any intermediate text shows, see for example Deaton and Muellbauer (1980b, ch. 2), the Slutsky matrix must be symmetric and negative

semi-definite. The symmetry property is not readily turned into simple intuition; negativity implies that the diagonal elements of the matrix are non-positive, a proposition often referred to as 'the law of demand'. The four properties, adding-up, homogeneity, symmetry and negativity, essentially exhaust the implications of utility maximization, so that any empirical demand functions that satisfy them can be regarded as having been generated by utility maximization, or by rational choice, with 'rational' defined, following Gorman (1981), as 'having smooth strictly quasi-concave preferences, and being greedy'.

Stone (1954b) was the first to attempt to use this theory directly to confront the data. He started from a (general) linear expenditure system of the form

$$p_i q_i = \sum_j a_{ij} p_j + b_i x \tag{30}$$

where a_{ij} and b_i are unknown parameters. Stone showed that, in general, the system (30) does not satisfy the four requirements, but will do so if, and only if, the parameters are restricted so that the model can be written in the form

$$p_i q_i = p_i \gamma_i + \beta_i (x - p \cdot \gamma) \tag{31}$$

with the β-parameters summing to unity. In this form the model is known as the linear expenditure system. As Samuelson (1947–8) and Geary (1949–50) had earlier shown, the utility function corresponding to (31) has the form

$$u = \sum_i \sum \beta_i \ln(q_i - \gamma_i), \tag{32}$$

sometimes referred to (somewhat inappropriately) as the Stone–Geary utility function. It can be thought of as a sum of Bernoulli utility functions of the quantity of each good above the minimal γ's.

Stone's achievement lay not in deriving the demand functions, but in thinking to estimate them. The demand functions (30), even if fitted to the data by least-squares, require non-linear optimization, and Stone invented a simple and not very efficient scheme, but one that allowed him to obtain parameter estimates and a good fit to interwar British data for a six commodity disaggregation of expenditures. This was a major breakthrough, not only in demand analysis, but also in applied econometrics in general. Indeed, much of demand analysis for a decade or so after Stone's paper consisted of applying better algorithms and faster computers to the fitting of Stone's model to different data sets.

The linear expenditure system offers a demand model for a system of, say n goods, and requires only $2n - 1$ parameters, a degree of parsimony that was very important in allowing the model to be estimated on very short time-series data. However, such economy brings its own price, and the linear expenditure system is very restrictive in the sort of behaviour that it can allow. In particular, and pathological cases apart, the model cannot allow inferior goods (goods the demand for which falls as total outlay increases), nor can it allow goods to be complements rather than substitutes. (As defined by Hicks (1939) goods i and j are complements if the (i, j)th term in the Slutsky matrix is negative, so that the utility compensated cross-price response of i to an increase in the price of j is positive.) Normal (non–inferior) goods that are substitutes for one another may be the most important case, but they do not encompass everything that we might want to study. The linear expenditure system also implies that the marginal propensity to consume each good is the same no matter what is the total to be spent, and many cross-section studies of household budgets have suggested that this is not in fact the case.

Unfortunately, it is quite difficult to write down utility functions that will lead to more general demand functions than those of the linear expenditure system, nor is there any obvious way of generalizing Stone's procedure of writing down functions and making them consistent with the theory. Progress was only really made once applied demand analysis started using 'dual' formulations of preferences to specify demands. In the demand context, duality refers to a switch of variables, from quantities to prices, so that utility becomes a function, not directly of quantities consumed, but indirectly of prices and total expenditure. This indirect utility formulation is given by

the function $\psi(x, p)$, already used above, and this is simply the maximum attainable utility from total outlay x at prices p. Since $\psi(x, p) = u$, and the function is monotone increasing in x, it can be inverted to give $x = c(u, p)$, known as the 'cost function', since it gives the minimum necessary cost that is required to reach the utility level u. By a theorem usually attributed to Shephard (1953) and to Uzawa (1964), these two functions contain a complete representation of preferences; provided preferences are convex, and provided the functions satisfy homogeneity and convexity (or concavity) conditions, preferences can be reconstructed from knowledge of either of the two functions. It is also very easy to move from either cost or indirect utility functions to the demand functions. For the indirect utility function, we have Roy's identity (Roy 1943).

$$q = -\nabla_p \psi(x,p)/\psi_x(x,p) \equiv g(x,p) \qquad (33)$$

which immediately yields demand functions from preferences in a form that are suitable for estimation, while for the cost function, we have Shepard's Lemma (1953),

$$q = \nabla_p c(u,p) = \nabla_p c[\psi(x,p),p] \equiv g(x,p) \quad (34)$$

where, as in (33), the operator ∇ denotes a vector of partial derivatives.

Demand analysis now had a high road to specification. Think of some quasi-convex decreasing function of the ratios of price to total outlay and call it an indirect utility function, or think of some function of utility and prices that is increasing in its arguments and linearly homogeneous and concave in prices and call it a cost function. Either way, and with only simple differentiation, new (and sometimes) interesting demand functions will be generated. Alternatively, and even more importantly, it is possible to use theory to aid and check out empirical knowledge. If it is known that the marginal propensity to spend on food is a declining function of total expenditure, or if it is thought likely that some goods do not depend very directly on the prices of other goods, it is relatively straightforward to find out what preferences (if any) will yield the result. It becomes possible, not just to generate demand functions serendipitously, but to generate good and useful ones deliberately.

There are many examples that could be cited from the literature. One of the most widely used in the *translog* model which was first proposed in 1970 by Jorgenson and Lau, see Christensen et al. (1973) for a convenient reference. To derive the translog, write the indirect utility function in terms of the ratios of prices to outlay, $r = p/x$, and approximate the indirect utility function as a second order polynomial in the logarithms of r. Application of Roy's identity yields demand functions in which the budget share of each good is the ratio of two functions, each of which is linear in the logarithms of the price to outlay ratios. Estimation of these rational functions, like estimation of the linear expenditure system, requires the use of non-linear maximization techniques. A related model, the 'almost ideal demand system' (AIDS) has been proposed by Deaton and Muellbauer (1980a), and I use this to illustrate some of the issues that arise with the current generation of demand models. The AIDS is specified by the logarithm of its cost function which takes the form

$$\ln c(u,p) = \alpha_0 + \sum_k \alpha_k \ln p_k + 0.5 \sum_k \times \sum_m \gamma_{km} \ln p_k \ln p_m + u \exp\left\{\sum_k \beta_k \ln p_k\right\}, \qquad (35)$$

so that, applying Shephard's lemma and rearranging, we have demand functions

$$p_i q_i / x \equiv w_i = \alpha_i + \beta_i \ln(x/P) + \sum_j \gamma_{ij} \ln p_j \qquad (36)$$

where P is a linearly homogeneous price index, the form of which can readily be inferred from (35). The parameters of the model must satisfy certain restrictions if (35) is to be a proper (log) cost function, and (36) a proper system of demand functions. The matrix of γ-parameters can be

taken to be symmetric in (35), but must be so in (36), and its rows and columns must add to zero for the homogeneity and adding-up properties to be satisfied. The β-parameters can be positive or negative, with positive values indicating luxury goods, and negative values necessities. The main advantage of the AIDS model in time-series applications is that the price index P can typically be approximated by some known price index selected before estimation, so that the demand system is linear in its parameters. In consequence, it can be estimated by ordinary least squares on an equation by equation basis, at least if the symmetry of the γ-matrix is ignored. The homogeneity restrictions can be tested equation by equation using a t- or F-test, and while imposing or testing symmetry requires an iterative procedure, estimation can be done by straightforward iterated restricted generalized least-squares, see Barten (1979) or Deaton (1974a) for further discussion.

The results of estimating the AIDS model are sufficiently similar to those from other models and other studies, see e.g. Barten (1969), Deaton (1974a), Christensen et al. (1973), and many others, that perhaps they can be taken as representative. What typically seems to happen is that the homogeneity restrictions appear *not* to be satisfied, so that in the application of AIDS to British data, Deaton and Muellbauer found, for example, that the F-test for transport had a value of 172 compared with the 5 per cent critical value of 4.8. Results on symmetry from AIDS and other systems are more mixed, and it now seems clear that testing symmetry is not usually possible given the amount of data typically available in time series, or put more positively, that there is no convincing evidence against symmetry. The difficulty is that symmetry involves a set of restrictions *across* different equations, so that unlike homogeneity, which involves tests *within* each equation, exact, small sample tests are not available. Researchers have therefore fallen back on asymptotically valid tests, and it turns out that these work very badly for the usual sort of samples, especially when there are more than a very small number of goods in the demand system. The papers by Laitinen (1978) and Meisner (1979) first established the problem, see also Evans and Savin (1982) and Bera et al. (1981) for further evidence.

The AIDS model, like the translog and several others, e.g. Diewert's (1973) 'generalized Leontief' system, fall into the class of 'flexible functional forms'. This criterion of flexibility, first proposed by Diewert (1971), is an important guarantee that the model is sufficiently richly parametrized so as to allow estimation of what are thought to be the main parameters of interest, typically the total expenditure elasticities, and the matrix of own and cross-price elasticities. A 'second order' flexible functional form is one that has sufficient parameters, so configured, that it is possible to set the value of the function, and of its first and second partial derivatives to any arbitrary set of (theoretically permissible) values. By applying Roy's identity or Shephard's lemma, it is clear that a cost or indirect utility function that is a second order flexible functional form will yield demand functions that are first-order flexible, so that it is possible for estimation to yield any set of price and expenditure elasticities that are consistent with utility theory. For empirical work, such a guarantee is important, because it ensures that the elasticities are being measured, not assumed. Contrast, for example, the linear expenditure system (31) with the AIDS model (36). Both could be fitted to the same set of data, and the parameter estimates of each could be used to generate a complete set of expenditure and price elasticities. But the linear expenditure system is *not* a flexible functional form, and so its estimated elasticities are not independent of one another, as is apparent from the fact that there are $2n - 1$ parameters compared with the total number of potentially independent elasticities, which is $(n - 1)(1 + n/2)$. (There are $n - 1$ independent demand equations, each of which has an expenditure elasticity, and n price elasticities; however, one price elasticity per equation is lost to homogeneity, and symmetry imposes a further $(n - 1)(n - 2)/2$ constraints.) The linear expenditure system does not therefore *measure* all the price and income elasticities, but determines them by a mixture of measurement and assumption, the main assumption being that of additive preferences, see Deaton (1974b) for further details. The AIDS, by contrast, has exactly

the right number of parameters to allow for intercepts and a full set of elasticities, so that when it (or the translog, or the generalized Leontief) is estimated, so is the full set of elasticities.

Being able to do this is a great step forward in methodology, but just as the linear expenditure system probably asks too little of modern data, (although not of the data available to Stone and the early pioneers of the systems approach), the second-order flexible functional forms probably ask too much, or equivalently, put too little structure on the problem. The consequences show up in large standard errors, a high frequency of apparently chance correlations, and a lack of robustness to functional form changes within the class of flexible functional forms, in other words, in all the standard symptoms of over-parametrization. These problems are particularly acute for the measurement of *price* elasticities, because in most time-series data, commodity prices tend to move together with relatively little variation in relative prices. And although the focus of most research on demand analysis over the last thirty years has been on the estimation and testing of price responses, there is certainly no consensus on what numbers, if any, are correct. Estimates obtained from the linear expenditure system are not credible because they are forced to satisfy an implausibly restrictive structure, while those from flexible functional forms are not credible because the data are not informative enough to supplement the lack of prior structure. Some intermediate forms are clearly required.

One of the attractions of flexible functional forms is their ability to approximate quite general forms for preferences. However, the models so far considered offer only approximations, and there is no guarantee that they have satisfactory *global* properties. Partly this is the standard problem that a fitted model will be forced to give a reasonable account of the data over the sample used for estimation, but may predict very badly elsewhere. But there are other deeper issues. Taking the AIDS as an example, estimation of (36) subject to symmetry and homogeneity will produce a system of estimated demand functions that will satisfy adding-up, homogeneity and symmetry for *all* values of x and p. However, there are two other important properties that are not assured. First, there is no guarantee that the predicted budget shares will necessarily lie between zero and one, so that there may be regions of price space in which the estimated model yields nonsensical predictions. Second, there is no way that the AIDS can be guaranteed to have a negative semi-definite Slutsky matrix for all prices, at least not without restricting parameters to the point where the model ceases to be a flexible functional form. The parameters could be chosen so as to satisfy negativity for some particular combination of prices and outlay, but there will be no guarantee that the law of demand will be satisfied elsewhere. In the translog model, it is possible to impose a restriction that guarantees negativity everywhere, but the model with the restriction has the property that all estimated own price elasticities must be less than minus one, independently of whether this is in fact true, and it almost certainly is not, see Diewert and Wales (1987). A demand system is described as 'regular' if it has a negative definite Slutsky matrix and predicts positive demands, and several empirical studies, see e.g. Wales (1977) for one of the first, found that estimated flexible functional forms were not regular over disturbingly large regions of even the parameter space used to estimate them. Caves and Christensen (1980), and later Barnett and Lee (1985) and Barnett et al. (1985), investigated the same problem theoretically by taking a known utility function, choosing the parameters of flexible functional forms to match its level and derivatives at a point, and then mapping out the regions of price space in which the systems remained regular. The results at least for the translog and the generalized Leontief model, were not good.

These regularity issues may seem of limited importance in practice, but this is far from being the case. One of the major reasons for being interested in complete empirical demand systems is to be able to examine the consequences of price changes, particularly of price changes that follow changes in government policy. The United States relies relatively little on indirect taxation as a source of public finance, but such is not the case in most of Europe, and the vast majority of developing countries maintain complex systems of

price wedges, particularly for foods and for agricultural production. The effects of such systems cannot be predicted without good information on how demands respond to price changes, nor can reforms be intelligently discussed. However, estimated demand systems that are not regular are not a great deal of help. All of the theory of welfare economics, of consumer surplus, of optimal taxation and of tax reform, *assumes* that demand behaviour is generated by utility maximization at the individual level, and implementation without regularity risks internal contradiction. For example, if compensated demand functions slope *upwards,* the government can generate a deadweight gain by imposing a distortionary tax. Of course, it may not be the empirical work that is wrong, but the theory that we used to try to model behaviour. If so, the estimated demand functions are still not useful, since we now have no idea what to do with them. But I doubt that evidence goes so far; it is not that behaviour itself is irregular, but that we have not yet found a good modelling strategy that contains a reasonable amount of prior information to supplement the paucity of data, and at the same time can deliver global regularity if it is warranted by the evidence.

A number of interesting experiments are currently under way that involve new modelling techniques. One possibility is that the Taylor series expansions that motivate most flexible functional forms are themselves inadequate to the task. In particular, Taylor approximations lose their ability to approximate if they are also asked to possess other properties of the functions that they are approximating. For example, we might want to test whether or not preferences are additively separable, as in the linear expenditure system. One strategy would be to write down some second-order approximation to preferences, estimate the resulting demand model, and then test whether or not the conditions imposed on the demands by additivity are satisfied. But this will not work in general, because there may be no additive system of demand equations that has the precise functional form demanded by the approximation. The same phenomenon is well illustrated by Stone's derivation of the linear expenditure system itself. The original general linear expenditure Eq. (29) can clearly be justified as a Taylor approximation to any set of homogeneous demand functions, and yet the imposition of only *symmetry* generates the demand system (30) which comes from the *additive* utility function (31). Additivity is not imposed, but linear expenditure systems are only symmetric if they are additive. Similarly many flexible function forms are only globally regular if they are homothetic, see for example, Blackorby et al. (1977). Several recent studies have proposed alternative ways of making functional approximations. Gallant (1982) has proposed using Fourier series approximations while Barnett (1983) has suggested that Laurent series can be used to generate demand models with good properties. Gallant's models are even more heavily parametrized than standard flexible functional forms, and there must be some question as to the suitability of trigonometrical functions for demand functions. Barnett's 'miniflex Laurent' model does not use the full flexibility of the Laurent series, but appears to have quite good approximation and regularity properties in practice, see Barnett and Lee (1985) and Barnett et al. (1985); even so, its estimation is complex, and many of the parameters have to be estimated subject to inequality constraints.

A second line of current research has abandoned the standard approach of econometric analysis, taking instead a completely non-parametric approach. Since many of the difficulties discussed above arise from choice of functional form, it is useful to ask how far it is possible to go without assuming any functional form at all. We know from standard revealed preference theory that two observed vectors of prices and quantities can be inconsistent with utility maximization; if bundle one is chosen when bundle two is available, so that bundle one is revealed preferred to bundle two, then no subsequent choice should reveal bundle two to be preferred to bundle one. Before embarking on the exercise of fitting some specific utility function to any finite collection of price and quantity pairs, one might then ask whether the collection is conceivably consistent with any set of preferences. If it is, then contradictions between an estimated system and the theory must be a matter of inappropriate functional form. The

conditions for utility consistency of a finite set of data were originally derived by Afriat (1967), who proposed a condition called cyclical consistency. Much later Varian (1982) not only provided an accessible and clear account of Afriat's results, but also recast the cyclical consistency condition into a 'generalized axiom of revealed preference (GARP)' that runs as follows. A bundle q^i is strictly directly revealed preferred to a bundle q if $p^i q^i > p^i q$, while q^i is revealed preferred to q, if there exists a sequence, $j, k, \ldots, m$ such that $p^i q^i \geq p^i q^j$, $p^j q^j \geq p^j q^k, \ldots, p^m q^m \geq p^m q$, so that q^i is directly or indirectly (weakly) revealed preferred to q. GARP is satisfied if for all q^i revealed preferred to q^j, it is not true that q^j is strictly directly revealed preferred to q^i, and given GARP the data can be rationalized by a continuous, strictly concave, and non-satiated utility function. Differentiability can also be ensured by a sight strengthening of GARP, see Chiappori and Rochet (1987). GARP is readily tested for any given set of data by checking the pairwise inequalities and using a simple algorithm provided by Varian to map out the patterns of indirect revealed preference. Repeated applications of the method to time-series data have nearly always confirmed the consistency of the data with the theory. In retrospect, it is clear that violations of GARP cannot occur unless some budget lines intersect, so that if, over time, economic growth has resulted in the aggregate budget line moving steadily outward with little change in slopes, GARP is bound to be satisfied. (However, post-war United States data budget planes do occasionally intersect, and Bronars (1987) has recently shown that hypothetical demands generated by selecting random points on the actual budget lines would more often than not fail GARP.)

The contradictions between the parametric and non-parametric approaches can perhaps be resolved by thinking of the latter as a modelling technique that uses a very large number of parameters, so that the failure of the parametric models to fit theory to data can be thought of as failure to parametrize the models sufficiently richly. But I have already argued that these models already have too many parameters, and adding more would only exacerbate the already serious problems of measurement. For many purposes, the theory is only useful if it is capable of delivering a description of the data that is reasonably parsimonious. There is also something rather simple minded about non-parametric techniques that tends to be disguised by the sophisticated and elegant expositions that have been given them by Varian and others. Consider a very simple theory that says variable x should move directly with variable y as, for example, in the Euler Eq. (15) above which says that, under certainty consumption should grow from period t to $t+1$ if and only if the real interest rate from t to $t+1$ is greater than some fixed constant. A non-parametric test on a finite set of data would accept the theory if, in fact, x, and y always did move together, and reject it if x and y ever moved in opposite directions. That such testing procedures are widely employed in the press and by the uninformed public is no reason for treating them seriously in economics.

I have so far discussed the formulation and estimation of demand functions, meaning the relationships between quantities, outlay, and prices, and this has been the topic of most applied demand analysis over the last thirty years. However, there is an older tradition of demand analysis, in which the object of attention is household budget data, and this literature has recently been enjoying something of a revival. Since household budget data typically come from a cross-section of households over a short period of time, usually within a single year, prices are treated as common to all sample points, so that the focus of attention becomes the relationship between demand and outlay and the influence of household composition on the pattern of household expenditures. The oldest, and perhaps only law of economics, Engel's Law that the share of food in the budget declines as total outlay increases, comes from Engel's (1857, published 1895) study of Belgian working-class families, and early empirical studies of demand were almost inevitably based on household surveys (see Stigler (1954) for a masterly review). The modern study of Engel curves, the relationships between expenditure and total outlay, begins (and almost ended) with Prais and Houthakker (1955). Prais and Houthakker studied

the shapes of Engel curves, the relationship between demand and households, particularly in relation to the choice of quality, a topic that has subsequently been unjustly neglected. The functional forms for Engel curves that Prais and Houthakker examined became the staple menu for most subsequent studies, even though only one of their forms, the linear Engel curve, is capable of satisfying adding-up, and the linear form typically performs very badly on the data. Since 1955 a number of other Engel curves have been proposed, notably the lognormal Engel curve of Aitchison and Brown (1957), and Leser's (1963) revival of the form suggested much earlier by Holbrook Working (1943). Working's form, which apparently escaped the attention of Prais and Houthakker, makes the budget share of each commodity a linear function of the logarithm of total outlay. The formulation is particularly useful, for not only is it capable of accounting for most of the curvature that is discovered in empirical Engel curves, but it is also consistent with utility theory, and corresponds to the case where the welfare elasticity of the cost of living is independent of income. Gorman (1981) has provided a general characterization theorem for Engel curves of the form

$$p_i q_i = \sum_k a_{ik}(p)\xi_k(x) \tag{37}$$

and has shown that the $\xi_k(\cdot)$ functions can be powers of x (polynomial Engel curves), or x multiplied by powers of log x (Engel curves relating budget shares to powers of the logarithm of outlay), or have trigonometric forms. This last form includes Fourier representations of Engel curves, while the first two allow Taylor or Laurent expansions for the expenditure/outlay and for the share/log-outlay forms. The Working–Engel curve is the first member of Gorman's 'share to log' class, and the theorem tells us that we may add quadratic or higher order terms to improve the fit. However, Gorman's paper contains a remarkable result; the matrix of the a-coefficients in (37) has rank at most equal to three. In consequence, the share to log and log-squared Engel curves are as general as any, as are the Engel curves of the quadratic expenditure system, see Howe et al. (1979). Given Gorman's results, and the empirical success of the Working form, it and its quadratic generalization deserve wide use in the analysis of budget studies. There is also accumulating evidence that such forms are indeed necessary. Thomas (1986), in a wide-ranging examination of household survey data from developing countries, has shown that Engel's Law itself does not appear to hold among the very poor, so that, in many cases, the share of the budget devoted to food at first rises with total outlay before falling in conformity with the Law.

Prais and Houthakker also proposed a much-used formulation for the effects of household composition on behaviour. It can be written

$$p_i q_i / m_i(a) = f_i\{x/m_0(a)\} \tag{38}$$

where a is a vector of household demographic characteristics (perhaps a list of numbers of people in each age and sex category) and m_i and m_0 are scalar valued functions known as the 'specific' and 'general scales' respectively. In this literature, scales are devices that convert family structure into numbers of equivalent adults, so that a family of two adults and two children might be two equivalent adults for theatre entertainment, three equivalent adults for food, and six equivalent adults for milk. The general scale is supposed to reflect the overall number of equivalent adults, so that the Prais and Houthakker model is a simple generalization of the idea that *per capita* demand should be a function of *per capita* outlay. Barten (1964), in a very important paper, took up the Prais–Houthakker idea of specific scales, but assumed that the arguments of the household utility function were the household consumption levels each deflated by the corresponding specific scale. The consequences of Barten's formulation are similar to those of Prais and Houthakker, but embody the additional insight that changes in family composition affect the effective shadow prices of goods, so that demographic changes will exercise, not only income, but also substitution effects on the pattern of demand. The story is often summarized by the phrase, 'if you have a wife and child, a penny bun costs three-pence',

quoted in Gorman (1976), but the really far-reaching substitution effects of children are probably on time use and labour supply, particularly of women.

Since household surveys typically contain large samples of households, there is less need for theory to save degrees of freedom, and it is possible to estimate quite general functional forms that link expenditures to household composition patterns and then to interpret the results in terms of the various models. In addition, neither the Prais–Houthakker nor the Barten model seem to yield easily implemented functional forms, e.g. linear ones, nor is it clear that either model is even identified on a single cross-sectional household survey in which all prices are constant, see for example Muellbauer (1980) and Deaton (1986a). However, some empirical results for the two models can be found in Muellbauer (1977, 1980) and in Pollak and Wales (1980, 1981) who also examine Gorman's (1976) extension of Barten's model in which additional people are supposed to bring with them fixed needs for particular commodities. The fixed needs model is close to the formulation proposed by Rothbarth (1943) for measuring the costs of children. Rothbarth pointed out that there are certain commodities, adult goods, that are not consumed by children, so that when children are added to a household, the only effects on the household's consumption of adult goods will be the income effects that reflect the fact that, with unchanged total resources, the household is now poorer. Deaton et al. (1985) have recently attempted to test Rothbarth's contention, and in their Spanish data it seems possible to identify a sensible group of adult goods, the expenditure on each of which changes with additional children in the same way as they change in response to changes in outlay.

Studies of the effects of family composition on household expenditure patterns have frequently been concerned, not only with estimating demands, but also with attempts to measure the 'cost' of children. It would take me too far afield to do justice to this topic here. Readers interested in this controversial area should perhaps start with Rothbarth (1943), who in a few pages makes a very simple and quite convincing case, and look also at Nicholson (1976). Pollak and Wales (1979) weigh in on the opposite side, and claim that it is impossible to measure child costs from expenditure data. My own position is argued in Deaton and Muellbauer (1986); there are certainly grave problems to be overcome in moving from the analysis of household survey data to the measurement of the costs of children, and it is clear that identifying assumptions must be made that are more severe and more controversial than those required, for example, to go from demand functions to consumer surplus. But that does not mean that it is not possible for such assumptions to be proposed and to be sensibly discussed.

See Also

- ▶ Bequests and the Life Cycle Model
- ▶ Consumer Expenditure (New Developments and the State of Research)
- ▶ Demand Theory
- ▶ Euler Equations
- ▶ Friedman, Milton (1912–2006)
- ▶ Rational Expectations

Bibliography

Abowd, J.M., and D. Card. 1985. *The covariance structure of earnings and hours changes in three panel data sets*. Princeton University, mimeo.

Afriat, S.N. 1967. The construction of a utility function from expenditure data. *International Economic Review* 8: 67–77.

Aitchison, J., and J.A.C. Brown. 1957. *The lognormal distribution*. Cambridge: Cambridge University Press.

Altonji, J. 1986. Intertemporal substitution in labor supply: Evidence from micro data. *Journal of Political Economy* 94: S176–S215.

Altonji, J., and A. Siow. 1985. Testing the response of consumption to income changes with (noisy) panel data. Industrial Relations Section Working Paper No.186. Princeton University, mimeo.

Ando, A., and F. Modigliani. 1963. The life-cycle hypothesis of saving: Aggregate implications and tests. *American Economic Review* 53: 55–84.

Ashenfelter, O. 1984. *Macroeconomic analyses and microeconomic analyses of labor supply*. Presented to Carnegie-Rochester Conference, Bal Harbor, Florida, November 1983.

Ashenfelter, O., and J. Ham. 1979. Education, unemployment, and earnings. *Journal of Political Economy* 87: S99–116.

Barnett, W.A. 1983. New indices of money supply and the flexible Laurent demand system. *Journal of Business and Economic Statistics* 1: 7–23.

Barnett, W.A., and Y.W. Lee. 1985. The global properties of the miniflex Laurent, generalized Leontief, and translog flexible functional forms. *Econometrica* 53: 1421–1437.

Barnett, W.A., Y.W. Lee, and M. Wolfe. 1985. The three dimensional global properties of the miniflex Laurent, generalized Leontief, and translog flexible functional forms. *Journal of Econometrics*: 3–31.

Barten, A.P. 1964. Family composition, prices, and expenditure patterns. In *Econometric analysis for national economic planning*, ed. P.E. Hart, G. Mills, and J.K. Whitaker. London: Butterworth.

Barten, A.P. 1969. Maximum likelihood estimation of a complete system of demand equations. *European Economic Review* 1: 7–23.

Bean, C.R. 1985. The estimation of surprise models and the surprise consumption function. Centre for Economic Policy Research (London), Discussion Paper No.54, mimeo.

Bera, A.K., R. Byron, and C.M. Jarque. 1981. Further evidence on asymptotic tests for homogeneity in large demand systems. *Economics Letters* 8: 101–105.

Bernanke, B.S. 1984. Permanent income, liquidity, and expenditure on automobiles: Evidence from panel data. *Quarterly Journal of Economics* 99: 587–614.

Bernheim, B.D., A. Schleiffer, and L.H. Summers. 1985. Bequests as a means of payment. *Journal of Political Economy*: 1045–1076.

Blackorby, C., D. Primont, and R.R. Russell. 1977. On testing separability restrictions with flexible functional forms. *Journal of Econometrics* 5: 195–209.

Blinder, A.S. 1975. Distribution effects and the aggregate consumption function. *Journal of Political Economy* 83: 447–475.

Blinder, A.S., and A.S. Deaton. 1985. The time series consumption function revisited. *Brookings Papers on Economic Activity* 2: 465–511.

Breeden, D. 1979. An intertemporal asset pricing model with stochastic consumption and investment opportunities. *Journal of Financial Economics* 7: 265–296.

Bronars, S.G. 1987. The power of non-parametric tests of preference maximization. *Econometrica* 55: 693–698.

Browning, M.J., A.S. Deaton, and M.J. Irish. 1985. A profitable approach to labor supply and commodity demands over the life cycle. *Econometrica* 53: 503–543.

Campbell, J.Y. 1987. Does saving anticipate declining labor income? An alternative test of the permanent income hypothesis. *Econometrica* 55: 1249–1273.

Campbell, J.Y., and N.G. Mankiw. 1986. Are output fluctuations transitory? National Bureau of Economic Research Working Paper 1916, processed.

Caves, D.W., and L.R. Christensen. 1980. Global properties of flexible functional forms. *American Economic Review* 70: 422–432.

Chiappori, P.-A., and J.-C. Rochet. 1987. Revealed preferences and differentiable demand. *Econometrica* 55: 687–691.

Christensen, L.R., D.W. Jorgenson, and L.J. Lau. 1973. Transcendental logarithmic production frontiers. *Review of Economics and Statistics* 55: 28–45.

Cochrane, J.H. 1986. How big is the random walk in GNP? Department of Economics, University of Chicago, processed.

Davenant, C. 1699. *Essay upon the probable methods of making a people gainers in the balance of trade*. London.

Davidson, J.E.H., et al. 1978. Econometric modelling of the aggregate time-series relationship between consumers expenditure and income in the United Kingdom. *Economic Journal* 88: 661–692.

Davies, J.B. 1980. Uncertain lifetime, consumption and dissaving in retirement. *Journal of Political Economy* 89: 561–577.

Deaton, A.S. 1974a. The analysis of consumer demand in the United Kingdom, 1900–1970. *Econometrica* 42: 341–367.

Deaton, A.S. 1974b. A reconsideration of the empirical implications of additive preferences. *Economic Journal* 84: 338–348.

Deaton, A.S. 1975. The structure of demand in Europe 1920–1970. In *The Fontana economic history of Europe*, ed. C.M. Cippola, vol. 5. London: Collins-Fontana.

Deaton, A.S. 1977. Involuntary saving through unanticipated inflation. *American Economic Review* 67: 899–910.

Deaton, A.S. 1986a. Demand analysis. In *Handbook of econometrics*, ed. Z. Griliches and M. Intriligator, vol. 3. Amsterdam: North-Holland.

Deaton, A.S. 1986b. Life-cycle models of consumption: Is the evidence consistent with the theory? NBER Working Paper No.1910, processed.

Deaton, A.S., and J. Muellbauer. 1980a. An almost ideal demand system. *American Economic Review* 70: 312–326.

Deaton, A.S., and J. Muellbauer. 1980b. *Economics and consumer behaviour*. New York: Cambridge University Press.

Deaton, A.S., and J. Muellbauer. 1986. On measuring child costs, with applications to poor countries. *Journal of Political Economy* 94: 720–744.

Deaton, A.S., J. Ruiz-Castillo, and D. Thomas. 1985. The influence of household composition on household expenditure patterns: Theory and Spanish evidence. Woodrow Wilson School, Princeton University, processed.

Dickey, D.A., and W.A. Fuller. 1981. Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica* 49: 1057–1072.

Diewert, W.E. 1971. An application of the Shephard duality theorem: A generalized Leontief production function. *Journal of Political Economy* 79: 481–507.

Diewert, W.E. 1973. Functional forms for profit and transformation functions. *Journal of Economic Theory* 6: 284–316.

Diewert, W.E. 1974. Intertemporal consumer theory and the demand for durables. *Econometrica* 42: 497–516.

Diewert, W.E., and T.J. Wales. 1987. Flexible functional forms and global curvature conditions. *Econometrica* 55: 43–68.

Dixit, A.K. 1976. *Optimization in economic theory.* Oxford: Oxford University Press.

Dolde, W., and J. Tobin. 1971. Monetary and fiscal effects on consumption in consumer spending and monetary policy: The linkages. Boston: Federal Reserve Bank of Boston, Conference Series no. 5.

Duncan, G.J., and D.H. Hill. 1985. An investigation of the extent and consequences of measurement error in labor economic survey data. *Journal of Labor Economics*: 508–532.

Durlauf, S.N., and P.C.B. Phillips. 1986. Trends versus random walks in time-series analysis. Cowles Foundation Discussion Paper No.788. Yale University, New Haven, processed.

Eichenbaum, M.S., L.P. Hansen, and K. Singleton 1984. A time series analysis of representative agent models of consumption and leisure choice under uncertainty. Graduate School of Industrial Administration, Carnegie-Mellon University, Pittsburgh, mimeo.

Engel, E. 1895. Die lebenkosten Belgischer Arbeiter-Familien früher und jetzt. *International Statistical Institute Bulletin* 9: 1–74.

Epstein, L. 1975. A disaggregate analysis of consumer choice under uncertainty. *Econometrica* 43: 877–892.

Evans, M.K. 1967. The importance of wealth in the consumption function. *Journal of Political Economy* 75: 335–351.

Evans, G.B.A., and N.E. Savin. 1982. Conflict among the criteria revisited; the W, LR, and LM tests. *Econometrica* 50: 737–748.

Fisher, I. 1930. *The theory of interest*. New York: The Macmillan Company.

Flavin, M. 1981. The adjustment of consumption to changing expectations about future income. *Journal of Political Economy* 89: 974–1009.

Flemming, J.S. 1973. The consumption function when capital markets are imperfect: The permanent income hypothesis reconsidered. *Oxford Economic Papers* 25: 160–172.

Friedman, M. 1957. *A theory of the consumption function*. Princeton: Princeton University Press.

Gallant, A.R. 1982. Unbiased determination of production technologies. *Journal of Econometrics* 20: 285–323.

Geary, R.C. 1949–50. A note on 'a constant utility index of the cost of living'. *Review of Economic Studies* 18, 65–66.

Ghez, G., and G.S. Becker. 1975. *The allocation of time and goods over the life-cycle*. New York: Columbia University Press.

Goldberger, A.S. 1967. Functional form and utility: A review of consumer demand theory. Social Systems Research Institute, University of Wisconsin, processed.

Gorman, W.M. 1959. Separable utility and aggregation. *Econometrica* 27: 469–481.

Gorman, W.M. 1976. Tricks with utility functions. In *Essays in economic analysis*, ed. M. Artis and A.-R. Nobay. Cambridge: Cambridge University Press.

Gorman, W.M. 1981. Some Engel curves. In *Essays in the theory and measurement of consumer behaviour in honour of Sir Richard Stone*, ed. A.S. Deaton. Cambridge: Cambridge University Press.

Gorman, W.M. 1982. Facing an uncertain future. IMSS Technical Report No.359, Stanford University, processed.

Grossman, S.J., and R.J. Shiller. 1981. The determinants of the variability of stock market prices. *American Economic Review, Papers and Proceedings* 71: 222–227.

Hall, R.E. 1978. Stochastic implications of the life cycle-permanent income hypothesis: Theory and evidence. *Journal of Political Economy* 86: 971–987.

Hall, R.E., and F.S. Mishkin. 1982. The sensitivity of consumption to transitory income: Estimates from panel data on households. *Econometrica* 50: 461–481.

Hansen, L.P., and K.J. Singleton. 1982. Generalized instrumental variables estimation of non-linear rational expectations models. *Econometrica* 50: 1269–1286.

Hayashi, F. 1982. The permanent income hypothesis: Estimation and testing by instrumental variables. *Journal of Political Economy* 90: 895–916.

Hayashi, F. 1985a. Permanent income hypothesis and consumption durability: Analysis based on Japanese panel data. *Quarterly Journal of Economics*: 183–206c.

Hayashi, F. 1985b. Tests for liquidity constraints: A critical survey. Osaka University and NBER, processed. Presented at the Fifth World Congress of the Econometric Society, Cambridge, MA, August 1985.

Heckman, J.J. 1971. Three essays on the supply of labor and the demand for goods. Unpublished PhD thesis, Princeton University.

Hicks, J.R. 1939. *Value and capital*. Oxford: Oxford University Press.

Houthakker, H.S. 1961. An international comparison of personal saving. *Bulletin of the International Statistical Institute* 38: 55–70.

Houthakker, H.S. 1965. On some determinants of saving in developed and underdeveloped countries. In *Problems in Economic Development*, ed. A.G. Robinson. London: Macmillan.

Houthakker, H.S., and L.D. Taylor. 1970. *Consumer demand in the United States: Analysis and projections*. 2nd ed. Cambridge, MA: Harvard University Press.

C

Howe, H., R.A. Pollak, and T.J. Wales. 1979. Theory and time series estimation of the quadratic expenditure system. *Econometrica* 47: 1231–1247.

Koskela, E., and M. Viren. 1982a. Saving and inflation: Some international evidence. *Economics Letters* 9: 337–344.

Koskela, E., and M. Viren. 1982b. Inflation and savings: Testing Deaton's hypothesis. *Applied Economics* 14: 579–590.

Kotlikoff, L.J., and L.H. Summers. 1981. The role of intergenerational transfers in aggregate capital accumulation. *Journal of Political Economy* 89: 706–732.

Kuznets, S. 1946. *National income: A summary of findings.* National Bureau of Economic Research. New York: Arno Press.

Kuznets, S. 1962. Quantitative aspects of the economic growth of nations: VII: The share and structure of consumption. *Economic Development and Cultural Change* 10: 1–92.

Kydland, F.E., and E.C. Prescott. 1982. Time to build and aggregate fluctuations. *Econometrica* 50: 1345–1370.

Laitinen, K. 1978. Why is demand homogeneity so often rejected? *Economics Letters* 1: 187–191.

Leff, N. 1969. Dependency rates and saving rates. *Economic Journal* 59: 886–896.

Leser, C.E.V. 1963. Forms of Engel functions. *Econometrica* 31: 694–703.

Lucas, R.E. 1976. Econometric policy evaluation: A critique. In *The Phillips curve and labor markets*, Carnegie-Rochester Conference Series on Public Policy 1, ed. K. Brunner and A. Meltzer. Amsterdam: North-Holland.

Lucas, R.E. 1981. Introduction. In *Studies in business cycle theory*, ed. R.E. Lucas. Cambridge, MA: MIT Press.

Lucas, R.E., and L. Rapping. 1969. Real wages, employment, and inflation. *Journal of Political Economy* 77: 721–754.

MaCurdy, T.E. 1981. An empirical model of labor supply in a life-cycle setting. *Journal of Political Economy* 89: 1059–1085.

Mankiw, N.G., and M. Shapiro. 1985. Trends, random walks, and tests of the permanent income hypothesis. *Journal of Monetary Economics* 16: 165–174.

Mankiw, N.G., and M. Shapiro. 1986. Do we reject too often? Small sample properties of tests of rational expectations models. *Economics Letters* 20: 139–145.

Mankiw, N.G., J.J. Rotemberg, and L.H. Summers. 1985. Intertemporal substitution in macroeconomics. *Quarterly Journal of Economics* 100: 225–251.

Meisner, J.F. 1979. The sad fate of the asymptotic Slutsky symmetry test. *Economics Letters* 2: 231–233.

Mirer, T.W. 1979. The wealth–age relationship among the aged. *American Economic Review* 69: 435–443.

Modigliani, F. 1970. The life-cycle hypothesis of saving and inter-country differences in the saving ratio. In *Induction, growth and trade: Essays in honour of Sir Roy Harrod*, ed. W.A. Eltis et al. Oxford: Clarendon Press.

Modigliani, F. 1986. Life cycle, individual thrift, and the wealth of nations. *American Economic Review* 76: 297–313.

Modigliani, F., and A. Ando. 1957. Tests of the life-cycle hypothesis of savings. *Bulletin of the Oxford Institute of Economics and Statistics* 19: 99–124.

Modigliani, F., and R. Brumberg. 1954. Utility analysis and the consumption function: An interpretation of cross-section data. In *Post-Keynesian economics*, ed. K.K. Kurihara. New Brunswick: Rutgers University Press.

Modigliani, F., and Brumberg. 1979. Utility analysis and aggregate consumption functions: An attempt at integration. In *The collected papers of Franco Modigliani*, ed. A. Abel, vol. 2. Cambridge, MA: MIT Press.

Muellbauer, J. 1977. Testing the Barten model of household composition effects and the cost of children. *Economic Journal* 87: 460–487.

Muellbauer, J. 1980. The estimation of the Prais–Houthakker model of equivalence scales. *Econometrica* 48: 153–176.

Muellbauer, J. 1985. *Habits, rationality and the life-cycle consumption function*. Nuffield College, Oxford, mimeo.

Nicholson, J.L. 1976. Appraisal of different methods of estimating equivalence scales and their results. *Review of Income and Wealth* 22: 1–11.

Phillips, P.C.B., and P. Perron. 1986. Testing for a unit root in a time series regression. Cowles Foundation Discussion Paper No.795, Yale University, New Haven, processed.

Pollak, R.A., and T.J. Wales. 1979. Welfare comparisons and equivalent scales. *American Economic Review* 69: 216–221.

Pollak, R.A., and T.J. Wales. 1980. Comparisons of the quadratic expenditure system and translog demand system with alternative specifications of demographic effects. *Econometrica* 48: 595–612.

Pollak, R.A., and T.J. Wales. 1981. Demographic variables in demand analysis. *Econometrica* 49: 1533–1551.

Prais, S.J., and H.S. Houthakker. 1955. *The analysis of family budgets*. Cambridge: Cambridge University Press.

Ramsey, F.P. 1928. A mathematical theory of saving. *Economic Journal* 38: 543–559.

Rothbarth, E. 1943. Note on a method of determining equivalent income for families of different composition. Appendix 4 in *War-time patterns of saving and spending*, ed. C. Madge, Occasional Paper 4, National Institute of Economic and Social Research, London.

Roy, R. 1943. *De l'utiliteé: contribution à la théorie des choix*. Paris: Herman.

Runkle, D.E. 1983. *Liquidity constraints and the permanent income hypothesis: Evidence from panel data*. MIT, processed.

Samuelson, P.A. 1947–8. Some implications of linearity. *Review of Economic Studies* 15: 88–90.

Shephard, R. 1953. *Cost and production functions*. Princeton: Princeton University Press.
Shorrocks, A.F. 1975. The age-wealth relationship: A cross-section and cohort analysis. *Review of Economics and Statistics* 57: 155–163.
Slutsky, E. 1915. Sulla teoria del bilancio del consumatore. *Giornale degli Economisti* 15: 1–26. English translation in *Readings in price theory,* ed. G.J. Stigler and K. Boulding. Chicago: Chicago University Press, 1952.
Smith, J.P. 1977. Family labor supply over the life cycle. *Explorations in Economic Research* 4: 205–276.
Spinnewyn, F. 1979a. Rational habit formation. *European Economic Review* 15: 91–109.
Spinnewyn, F. 1979b. The cost of consumption and wealth in a model with habit formation. *Economics Letters* 2: 145–148.
Stigler, G.J. 1954. The early history of empirical studies of consumer behavior. *Journal of Political Economy* 62: 95–113.
Stone, J.R.N. 1954a. *The measurement of consumers' expenditure and behaviour in the United Kingdom, 1920–1938*, vol. 1. Cambridge: Cambridge University Press.
Stone, J.R.N. 1954b. Linear expenditure systems and demand analysis: An application to the pattern of British demand. *Economic Journal* 64: 511–527.
Stone, J.R.N. 1964. Private saving in Britain, past, present, and future. *The Manchester School* 32: 79–112.
Stone, J.R.N. 1966. Spending and saving in relation to income and wealth. *L'Industria*: 471–499.
Surrey, M.J.C. 1974. Saving, growth, and the consumption function. *Bulletin of the Oxford Institute of Statistics* 36: 125–142.
Thomas, D. 1986. Essays on the analysis of Engel curves in developing countries. PhD thesis, Princeton University.
Uzawa, H. 1964. Duality principles in the theory of cost and production. *International Economic Review* 5: 216–220.
Varian, H.R. 1978. *Microeconomic analysis*. 2nd ed, 1984. New York: Norton.
Varian, H.R. 1982. The non-parametric approach to demand analysis. *Econometrica* 50: 945–973.
Wales, T.J. 1977. On the flexibility of flexible functional forms: An empirical approach. *Journal of Econometrics* 5: 183–193.
Watson, M.W. 1986. Univariate detrending method with stochastic trends. *Journal of Monetary Economics* 18: 49–75.
Working, H. 1943. Statistical laws of family expenditure. *Journal of the American Statistical Association* 38: 43–56.
Yaari, M.E. 1965. Uncertain lifetime, life insurance, and the theory of the consumer. *Review of Economic Studies* 32: 137–150.
Zeldes, S. 1985. Consumption and liquidity constraints: An empirical investigation. The Wharton School, University of Pennsylvania, processed.

Consumer Expenditure (New Developments and the State of Research)

Orazio P. Attanasio and Guglielmo Weber

Abstract

We provide an overview of recent developments in the life-cycle permanent income model under uncertainty, starting from the certainty equivalence case, and considering precautionary saving, the workings of insurance and credit markets, and non-standard preference structures.

Keywords

Aggregate consumption; Aggregate saving; Asset pricing; Asymmetric information; Buffer stocks; Certainty equivalence; Commitment; Consumer expenditure; Consumer expenditure (New developments and the state of research); Consumption function; Durable goods; Elasticity of intertemporal substitution; Equity premium puzzle; Euler equation; Excess sensitivity; Excess smoothness of consumption; Expected utility; Financial asset accumulation; Financial market imperfections; Habits; Impatience; Interest rate; Intertemporal optimization; Intertemporal prices; Labour supply; Lagrange multipliers; Life cycle hypothesis; Liquidity constraints; Marginal utility of consumption; Marginal utility of wealth; Myopia; Non-convexity; Perfect insurance; Permanent income life-cycle model; Power utility; Precautionary savings; Preferences; Private information; Prudence; Public finance; Quadratic utility; Quasi-hyperbolic discounting; Real business cycles; Reduced form equations; Retirement; Risk; Risk aversion; Saving; Social insurance; Social Security in the United States; Stone–Geary utility; Subjective discount rate; Time preference; Transaction costs; Uncertainty

JEL Classifications
E22

The state of research on consumer expenditure up to the mid-1980s is described in consumer expenditure. Here, we provide an overview of recent developments on the intertemporal model of consumer behaviour under uncertainty. We organize our discussion around what has been the workhorse model for the analysis of dynamic consumption behaviour – the life-cycle permanent income model. Although our discussion of the intertemporal model is self-contained, it is not meant to be an exhaustive survey of this large literature. We do not cover demand analysis, despite the many exciting developments that have occurred in recent years.

The permanent income life-cycle (PILC) model, introduced during the 1950s by Modigliani and Brumberg (1954) and Friedman (1957), still plays an important role in the consumption literature. The PILC model can be loosely defined as a framework where individuals maximize utility over time given a set of intertemporal trading opportunities. Consumption at different points in time is treated as different commodities, so that, given intertemporal trading opportunities, consumption in a given period depends on total (life-cycle) resources and (intertemporal) prices. Optimal consumption choices are such that the ratio of (expected) marginal utilities of consumption at different times equals the ratio of intertemporal prices. Therefore, the relationship between consumption and total resources is likely to depend on preferences (and in particular on the elasticity of intertemporal substitution and the rate at which the future is discounted) and on interest rates (as they represent intertemporal prices). If we allow for uncertainty, as we discuss below, risk will also enter as a potentially important determinant of consumption.

This model can generate implications and insights for many important questions not only in macroeconomics but also in public finance, and has therefore attracted much attention, both theoretically and empirically. Recent research has stressed the need to look at preferences on the one hand and markets on the other, as the policy implications are the result of both.

The Permanent Income Life-Cycle Model

In its simplest incarnation the PILC model considers a finite horizon, no uncertainty and very simple preferences. In such a situation, it is simple to translate the basic intuition of the model, to which we referred above, into a closed form solution for consumption that depends not just on current income but on the total amount of resources available to an individual and intertemporal prices. The problem of this specification, of course, is its lack of realism. Not only do consumers in reality face much more complicated intertemporal environments, but it is likely that these complications have a first-order effect on consumption choices. Therefore, the simplest version of the model is a useful way to convey the main ideas behind PILC, but it needs to be complicated considerably to be of use for policy analysis.

The introduction of uncertainty in the model, which makes it much more realistic, complicates the problem enormously. The first formalizations of the life-cycle model under uncertainty date back to the 1970s (Bewley 1977). Typically, one assumes that consumers maximize expected life-cycle utility choosing consumption and, in more general settings, leisure and financial asset holdings. Consumers are assumed to know the stochastic nature of their environment. Even with many simplifications on the nature of preferences, the model does not yield closed form solutions for consumption, except in the most special cases.

MaCurdy (1981, see also 1999) uses dynamic optimization techniques to derive necessary conditions for the optimal solution of the intertemporal optimization problem faced by consumers. The attractiveness of this approach lies in the fact that it cuts through the necessity of solving the model completely, which is a very hard task indeed, to focus on some useful implications of the model. In particular, these contributions focus on the basic first order condition, the so-called Euler equation, that equates the ratio of marginal utilities to intertemporal prices.

The first macro paper to take this approach is Hall (1978): under strong assumptions on preferences and returns, (non-durable) consumption is a random walk, that is:

$$E(C_{t+1}|I_t) = C_t \quad (1)$$

where I_t denotes information available at time t. This remarkable proposition requires that utility be quadratic in consumption (and additively separable over time, states of nature and in its other arguments, notably male and female leisure and durable goods). It also requires that there is at least one financial asset with fixed real return, and that this equals the time-preference parameter. If consumers have rational expectations, then:

$$C_{t+1} = C_t + \varepsilon_{t+1}E(\varepsilon_{t+1}|Z_t) = 0 \quad (2)$$

for all variables Z known at time t. A notable feature of Hall's model is that the Euler equation for consumption aggregates perfectly, because it involves linear transformations of the data. Hall used the Euler equation to test for the prediction implied by (2): no variable known to the consumer at time t should help predict the change in consumption between t and $(t + 1)$.

Hall's paper was the first of many contributions that exploited the Euler equation and the fact that such an approach does not require the complete specification of the environment in which the consumer lives, or even the complete budget constraint. Moreover, the approach is robust to the presence of various imperfections in some intertemporal markets. And while the specification with quadratic utility yields a linear equation for consumption, alternative specifications, with more plausible preferences, are easily introduced. For instance, in the case of power utility, an expression similar to (2) can be obtained for the log of consumption.

The price that one pays in using the Euler equations approach, which we discuss below, is that one does not obtain a closed form solution for consumption.

An approach that goes beyond the consideration of the Euler equations is taken up in an important paper by Flavin (1981).

Flavin (1981) adopts the same theoretical framework as Hall (1978), and assumes that no other asset is available to the consumer (as in Bewley 1977). However, Flavin develops a solution for consumption. To do so, she has to specify completely the stochastic environment in which the consumer lives and use particularly simple preferences. In particular, Flavin (1981) assumes that the only stochastic variable is labour income, that preferences are quadratic and that the consumer can save or borrow in a single asset with a fixed rate of interest. Under these conditions, Flavin shows that consumption is set equal to permanent income, and this is in turn defined as the present value of current and expected future incomes:

$$C_t = \frac{r}{1+r}A_t + \frac{r}{1+r}\sum_{k=0}^{\infty} E\left(y_{t+k}|I_t\right) \quad (3)$$

where A denotes financial wealth and y is labour income. In this model, the first difference in consumption equals the present value of income revisions, due to the accrual of new information between periods t and $(t + 1)$:

$$\Delta C_t = \frac{r}{1+r}\sum_{k=0}^{\infty}\frac{1}{(1+r)^k}\left[E\left(y_{t+k}|I_t\right) - E\left(y_{y+k}|I_{t-1}\right)\right]. \quad (4)$$

Equation (3) makes clear the main implications of the model: consumption depends on the present discounted value of future expected income. The interest rate plays the important role of converting future resources to present ones and therefore constitutes an important determinant of consumption. Flavin (1981) noticed that Eq. (3) imposes cross-equation restrictions on the joint time series process for income and consumption. A similar approach had been followed by Sargent (1978) and, subsequently, by Campbell (1987) who noticed that an implication of (4) is that saving predicts future changes in income, the so-called 'saving for a rainy day' motive.

One of the main implications of the PILC model, particularly evident in Eq. (3), is that, in appraising the effects of a given policy, for instance a tax reform that affects disposable income, a distinction must be drawn between permanent and temporary changes (Blinder and Deaton 1985; Poterba 1988).

Another feature of Flavin's model is that the closed form solution for consumption is the same under certainty and uncertainty, as long as expected values of future incomes are taken. This is a direct consequence of the assumption of quadratic utility that makes the marginal utility linear in consumption. For this reason, it is often referred to as the certainty equivalent model.

Extensions of the Simple Certainty Equivalent Model

The certainty equivalent model is appealing for its simplicity, but its implications are typically rejected by the data: Hall and Mishkin (1982) were particularly influential in suggesting that some of the model implications were rejected in micro data. At the same time, the model with quadratic preferences was perceived to be too restrictive in its treatment of financial decisions: quadratic preferences imply increasing absolute risk aversion in consumption (or wealth), something that is unappealing on theoretical grounds and strongly counterfactual (riskier portfolios are normally held by wealthier households). Quadratic preferences also imply that the willingness to substitute over time is a decreasing function of consumption: poor consumers should react much more to interest rate changes than rich consumers, after allowance has been made for the wealth/income effect.

The alternative adopted in much of the literature has been to assume power utility and to allow for the existence of a number of risky financial assets. Once one deviates from quadratic utility, however, and/or allows for stochastic interest rates, one loses the ability to obtain a closed form solution for consumption. Many of the studies that made this choice, therefore, focused on the study of the Euler equations derived from the maximization problem faced by the consumer. The basic first-order conditions used in this literature are two:

$$U_{ct} = \lambda t, \tag{5}$$

$$\lambda_t = E\left(\lambda_{t+1}\frac{1+r^k_{t+1}}{1+\delta}\,|\,I_t\right). \tag{6}$$

Equation (5) says that, at each point in time, the marginal utility of consumption equals the Lagrange multiplier associated with the budget constraint relevant for that period, which is sometimes referred to as the marginal utility of wealth. The second condition, Eq. (6), that is derived from intertemporal optimality, dictates the evolution of the marginal utility of wealth (δ is a subjective discount rate). An equation of this type has to hold for each asset k for which the consumer is not at a corner. This is because the consumer is exploiting that particular intertemporal margin.

The attractiveness of Euler equations is that one can be completely agnostic about the stochastic environment faced by the consumer, about the time horizon, about the presence of imperfections in financial markets (as long as there is at least one asset that the consumer can freely trade), about the presence of transaction costs in some component of consumption or labour supply. All relevant information is summarized in the level of the marginal utility of wealth. The approach is conceptually similar to the use of an (unobservable) fixed effect in econometrics. By taking first differences, one eliminates the unobservable marginal utility of wealth and is left only with the innovations to Eq. (6).

Early papers along these lines were Hansen and Singleton (1982, 1983), who used power utility (also known as isoelastic, isocurvature or CRRA) as it has more appealing theoretical properties (relative risk aversion is constant in wealth or consumption, the elasticity of intertemporal substitution is also a constant). If we substitute Eq. (5) into (6) and using the properties of the CRRA utility function, the Euler equations for consumption corresponding to each asset (k) will be:

$$E\left\{\left(\frac{C_{t+1}}{C_t}\right)^{-\gamma}\frac{1+r^k_{t+1}}{1+\delta}\right\} = 1 \tag{7}$$

where γ is a curvature parameter (equal to the relative risk aversion parameter and to the reciprocal of the elasticity of intertemporal substitution) and δ, the subjective discount rate, measures impatience.

An equation such as (7) can be log-linearized to obtain (see Hansen and Singleton 1983):

$$\Delta \log C_{t+1} = k + \frac{1}{\gamma} \log\left(1 + r^{k}_{t+1}\right) + \varepsilon_{t+1}. \quad (8)$$

Although consumption appears on the left-hand side of Eq. (8), this equation is not a consumption function, but an equilibrium condition. It cannot explain or predict consumption levels: consumption is crucially determined by the residual term ε_{t+1} and there is nothing that tells us what determines such a term or how this term changes with news about income, interest rates or any other relevant variable.

The Euler equation for a single asset can identify the elasticity of intertemporal substitution, a key parameter for the evaluation of the welfare costs of interest taxation (Boskin 1978; Summers 1981) and for the analysis of real business cycles (King and Plosser 1984). The joint estimation of several Euler equations can help identify the pure discount rate parameter (governing patience), but also shed light on risk aversion, given that different assets typically have different risk characteristics.

The derivation of a closed form solution for consumption when certainty equivalence does not hold was first successfully tackled by Caballero (1990, 1991). Caballero (1991) took the Flavin model with known finite life, and constant absolute risk aversion (CARA) preferences, and showed that, when the optimal consumption age profile is flat with no uncertainty, it is increasing with income uncertainty. This change in the slope of the consumption profile was described as precautionary saving, because early in life consumers save more if labour income is more uncertain. Later work by Gollier (1995) and Carroll and Kimball (1996) established that a similar result holds whenever the third derivative of the utility function is positive, a feature of preferences labelled prudence. Both CARA and power utility exhibit prudence. The presence and size of precautionary savings is a matter of great relevance for public policy, in so far as public insurance schemes covering such risks as unemployment, health and longevity should reduce the need for consumers to accumulate assets.

The great merit of the model with prudence is that it highlights the need to save for rainy days even if sunny days are equally likely. An increased variance in the shocks to income reduces consumption even if expected income does not change. In the case of discrete variables, such as unemployment or illness, changes in first and second moments occur simultaneously, but this is not the case for continuous variables. The ability to distinguish between first and second moment effects is of crucial importance in the analysis of public policy, because of its social insurance characteristics.

The solution of the Bewley model with more general utility functions has to be computed numerically or rely on approximations. Several studies in the early 1990s took up the challenge of characterizing such solutions. Deaton (1991) studied a model with power utility and infinite life. Deaton considered the existence of liquidity or no-borrowing constraints, and showed that impatient consumers would hold limited assets to insure against low income draws. Carroll (1992) instead covered the case of finite lives, and showed that, if consumers are sufficiently impatient and their labour income is subject to both permanent and temporary shocks, they set consumption close to income. The model with impatient consumers under labour income uncertainty has been labelled ‘the buffer stock model’, because saving is kept to the lowest level compatible with the need to buffer negative income shocks. Later work by Attanasio et al. (1999) and Gourinchas and Parker (2002) clarifies the role played by age-related changes in demographics and the hump-shape age profile of labour income in generating income tracking for relatively young consumers (micro data show that financial asset accumulation starts around age 40). Hubbard et al. (1994, 1995) show instead how precautionary motives interact with the insurance properties of Social Security in the United States.

Many of the papers cited in the preceding paragraph consider relatively simple versions of the life-cycle model. In particular, a single non-durable commodity is assumed and preferences are assumed to be additively separable

with leisure and over time. While this greatly simplifies the solution, the construction of a more realistic and complex model has become an important area of research. This development follows from the recognition that, for many purposes, and in particular for policy analysis, a model that delivers consumption as a function of exogenous variable is a very useful tool indeed.

This area of research has to deal with two important issues. First of all, the model can become very quickly, from a numerical point of view, very difficult to solve. The large number of state variables that characterize the solution of reasonably realistic models and the consideration of discrete choices and non-convexities linked to transaction costs can push the numerical capabilities of even very powerful computers. Second and even more importantly, if one wants to obtain solutions for consumption in a dynamic context, one has to characterize completely the stochastic environment in which the consumer lives. This contrast sharply with the Euler equation approach that allowed the researcher to be agnostic about most aspects of the environment and, under certain conditions, avoid solving difficult problems, such as labour supply, housing and other durable choices and so on. The Euler equation would hold regardless of the presence of non-convexities and other type of difficulties connected with these choices. These, instead have to be fully specified if one wants to work with a model that delivers a solution for consumption. These two difficulties constitute limits for the research in this area that, in all likelihood will not be overcome in the near future.

The Empirical Evidence on the PILC Model

Since its introduction in the 1950s, there is no consensus about the empirical relevance of the PILC model. While the model it is one of the main tools in modern macroeconomics and public finance, its empirical performance is mixed. In this section, we discuss two branches of the literature.

The life-cycle model with various sources of uncertainty and generic preferences generates decision rules and behaviour of great complexity. Consumption and saving choices depend in an unknown fashion on every single aspect of the stochastic environment faced by the consumer, for instance on the entire distribution of future wages and earnings opportunities, on pension arrangements, on the asset markets the consumer can access, on mortality risks and so on and so forth. The Euler equation approach allows researchers to deal in a rigorous fashion with extremely rich models and yet derive relatively simple implications to test some aspects of the model and, with the help of additional assumptions, to identify some of the structural parameters that inform individual behaviour. We now understand that Euler equations can be used to determine what type of preferences fits the available data and can therefore provide one of the building blocks (preferences) in the study of the questions above. We also know that the presence of liquidity constraints does not necessarily produce violations of Euler equations because, even when liquidity constraints are present, they might be rarely binding.

The Euler equation is robust to a number of market imperfections, but is silent about how consumption or its growth reacts to specific news about shocks, changes in interest rates, taxation and so on. It is therefore useless for specific policy analysis. In other words, while the parameters of an Euler equation can be estimated in a wide set of circumstances, and one can use the equation to test the specification of the model, none of these results will provide an answer to questions like what is the effect of a change in taxation or interest rates on the level of consumption and saving?

This important shortcoming of the Euler equation approach explains why such an approach, which has informed and dominated the large empirical literature on the validity of the life-cycle permanent income model is virtually absent in the public economics literature on, say, the effect of pension reforms on saving or on the effect of changes in the taxation of interest on saving. And yet the conceptual framework that is

behind the study of these issues is the same as that used to study consumption behaviour.

Policy analysis requires instead the availability of a consumption function, that is, a relation that explains consumption as a function of those variables that the consumer can take as exogenous at any given moment. Only in the simplest versions of the life-cycle model is it possible to derive an analytical expression for the consumption function. In general, given a set of assumptions on preference parameters and market and non-market opportunities, one has to rely on numerical solutions and/or approximations.

A less ambitious but potentially profitable approach that does not require numerical methods or incredibly rich data-sets is the estimation of reduced form equations, whose specification is informed by the life-cycle model. These are particularly useful in situations in which one analyses large (and possibly exogenous) changes to some of the likely determinants of consumption or saving. Such studies can address substantive issues and even test some aspects of the life-cycle model. Examples of studies of this kind include the reaction of consumption (and saving) to changes in pension entitlements (Attanasio and Brugiavini 2003; Attanasio and Rohwedder 2003; Miniaci and Weber 1999), to swings in the value of important wealth components (such as housing, Attanasio and Weber 1994) and to changes in specific taxes (Parker 1999; Souleles 1999; Shapiro and Slemrod 2003).

Below we review the empirical evidence on the PILC model, organizing it in two subsections. First we start with the empirical evidence derived from Euler equations. We then move on to evidence that considers the *levels* of consumptions, rather than its changes.

Evidence from Euler Equations

Two important empirical issues can be addressed with the study of Euler equations:

- What is the empirical relevance of the model? Is there a sensible specification of preferences that fits the observed data?
- What is the magnitude of the relevant preference parameters?

Tests of the Model

As mentioned above, a prediction of the model is that changes in consumption cannot be predicted by expected changes in income or any other variable known to the consumer at time $t-1$. This is the essence of the Hall (1978) test and of many others. Evidence that consumption can be predicted by lagged variables has been interpreted as indicative of liquidity constraints, myopic behaviour, misspecification of preferences and so on. The relationship between consumption and income has received considerable attention. The first to observe that the life-cycle model predicts no relation between the life-cycle profile of income and consumption was Thurow (1969). Thurow argued that the fact that consumption tracked income over the life-cycle was a rejection of the main implications of the PILC model. To this argument, essentially identical to many others proposed subsequently, Heckman (1974) replied that non-separability between consumption and leisure could explain such a relationship.

Despite this early exchange, after Hall (1978) a large fraction of the literature based on consumption Euler equations focused on the relationship between predictable changes in income and expected consumption growth. Hall and Mishkin (1982), as well as Campbell and Mankiw (1990, 1991) all report violations of this prediction, and label this finding 'excess sensitivity'. Excess sensitivity can be explained by the presence of liquidity constrained consumers, or of rule-of-thumb consumers, that is, consumers who let their expenditure track their income as a way to avoiding the complexities of choosing the optimal consumption path. However, consistently with Heckman's (1974) argument, excess sensitivity can be reconciled with the intertemporal optimization model if more general, and sensible, utility functions are used. In particular, if one assumes that leisure affects utility in a non-additive way, consumption changes respond to predictable labour income changes, whether or not leisure is a freely chosen variable. Finally, and importantly, the aggregation

issue proves to be important. Attanasio and Weber (1993) show that results obtained with improperly aggregated micro data are consistent with results obtained with aggregate data and indicate rejections of the model that instead disappear with properly aggregated data and rich enough preference structures.

To summarize the discussion so far, it seems that while simple tests of the life-cycle model seem to reject the implications from the model and in particular those derived from Euler equations, it is possible to find specification of preferences that do a good job at fitting the available data, especially for households that are headed by prime-aged individuals. Aspects that are crucial for fitting the data are the use of household level data, allowing for changes in consumption needs induced by changes in family composition and the use of preferences specifications that allow for the marginal utility of consumption to depend on labour supply.

Estimation of Preference Parameters

Recent research on consumption and saving has singled out three preference parameters for attention: the elasticity of intertemporal substitution, the relative risk aversion parameter and the subjective discount rate. The size of these parameters has important implications in many applications of the model, ranging from macroeconomics to public finance to financial economics.

Perhaps surprisingly, not much evidence has been accumulated on the discount factor from the estimation of Euler equations. This can be explained by the fact that in log-linearized versions of the Euler equation, the parameter is not identified, while non-linear versions of the model are ridden by a number of econometric problems, particularly in relatively small samples of the type used in Euler equation estimation (see Attanasio and Low 2004).

As for the distinction between the elasticity of intertemporal substitution (EIS) and the coefficient of risk aversion, it is absent in the most popular specifications used in the literature: a model where consumers maximize expected utility and preferences are iso-elastic and additively separable over time. In such a situation, the EIS is the reciprocal of the coefficient of relative risk aversion. Not many empirical papers have worked with preferences that allow for these two parameters to be disjoint.

An influential paper by Hall (1988) claimed that this parameter is close to zero. This finding has been challenged on various grounds. Attanasio and Weber (1993, 1995) point out that aggregation bias could be responsible for such a low estimate: they estimated a much higher elasticity (around 0.8) using UK and US cohort data (that is, data from repeated cross-sections, consistently aggregated over individuals born in the same years).

In the macro literature little attention has been paid to the possibility that the EIS may differ across consumers, particularly as a function of their consumption. A simple way to capture the notion that poor consumers may be less able to smooth consumption across periods and states of nature is to assume that the utility function does not depend on total (non-durable) consumption, but rather on the difference between consumption and needs. Thus we could retain the analytical attraction of power utility, but have $(C - C^*)$ as its argument, where C^* is an absolute minimum that the consumer must reach in each and every period. This functional form is known as Stone–Geary utility in demand analysis, and is the simplest way to introduce non-homotheticity in a demand system. One could interpret 'external habits' (Abel 1990; Campbell and Cochrane 1999) as a special way to parameterize C^* (by making it a fraction of past consumption of other consumers). Attanasio and Browning (1995), Blundell et al. (1994) and Atkenson and Ogaki (1996) are among the few examples of papers that explicitly allow for wealth-dependent EIS.

Demographics might also affect preferences, and might explain consumption changes and the shape of the consumption age profile, as argued by Attanasio et al. (1999) as well as Browning and Ejrnaes (2002).

Evidence from the Levels of Consumption

As stressed above, the Euler equation imposes some restrictions on the dynamics of consumption

but, on its own, does not determine the level of consumption. If one neglects numerical complications, a solution for consumption can be obtained by considering jointly the Euler equation and the sequence of budget constraints faced by the consumer as well as his or her initial wealth and a terminal condition. As noted by Sargent (1978), Flavin (1981) and later by Campbell (1987), the Euler equation and the intertemporal budget constraint imply a number of cross-equation restrictions for the joint time series processes of consumption and income. When one is able to obtain a closed form solution for consumption, as is the case with quadratic utility, these restrictions can be easily expressed in terms of a linear time series model, and tested.

Some of these restrictions are also implied by the Euler equation, while others are not. In particular, the restrictions on the contemporaneous correlation between income and consumption are not implied: as we stressed above, the Euler equation is silent about how news about income is translated into news about consumption.

Campbell and Deaton (1989) and West (1988) proposed a test that links the innovation to permanent income to consumption and presented evidence that aggregate consumption seems to be 'excessively smooth' in that it does not react enough to news about income. Campbell and Deaton make a connection between excess sensitivity and excess smoothness. Within the certainty equivalent model, they jointly model the consumption and income processes as a vector autoregression, assuming that income has a unit root plus some persistence. In this context, consumption changes reflect the permanent income innovation more than one-to-one: not only is the income shock permanent, but it also predicts future, smaller shocks of the same sign. This implies that over the business cycle consumption should be more volatile than income. But in actual aggregate data consumption is smoother than income: this is labelled 'excess smoothness', and is shown to be exactly equivalent to excess sensitivity.

Clearly the implications of a given set of intertemporal preferences for policy relevant questions depend crucially on the markets individuals have access to, on their imperfections and on the nature of the equilibrium they give rise to. The implications of complete markets would be very different from those one would derive if liquidity constraints or other markets imperfections were prevalent.

C

Insurance and Credit Markets

So far we have taken the assets the consumer can use to move resources over time as given and, in the simplest versions of the model, we have made very strong assumptions on this crucial aspect. For instance, we have assumed that consumers can borrow and lend at a fixed interest rate. The reality is, obviously, much more complex and, from a theoretical point of view, very many different environments have been studied. In particular, the possibilities open to a consumer depend on the market arrangements available. Below we discuss several of these market arrangements and briefly mention their implications for the determination of consumption.

Perfect Insurance Markets

If markets are complete and consumers can trade a full set of contingent claims without cost, individual risk will be completely diversified. In such a situation, a number of results deliver very useful predictions. In particular, it can be shown that a competitive equilibrium is symmetric and it is therefore possible to characterize the properties of competitive equilibria by considering the problem of a fictitious social planner, which, given a set of Pareto weights, maximizes social welfare. A strong implication of perfect markets is that the marginal utility of different consumers will move proportionally over time. The implication is very intuitive: the social planner faces a unique resource constraint, and marginal utility of all individuals, multiplied by the appropriate (and arbitrary) Pareto weight, will be equal to the multiplier associated to this unique constraint. As a consequence, marginal utility will move proportionally. If utility is isoelastic, consumption moves proportionally. These implications, stressed by Townsend (1994), have been tested in several

papers (Cochrane 1991; Attanasio and Davis 1996; Hayashi et al. 1996).

Many Assets

When there are many assets, one can derive an Euler equation such as (7) for each of the assets for which the consumer is not at a corner. The Euler equations for consumption with different assets naturally ties up with asset pricing equations. This approach to asset pricing was developed by Breeden (1979) and Lucas (1978), and extended to the case of non-additive separability of consumption and leisure in an incomplete markets setting by Bodie et al. (1992). The model we sketched above is quite restrictive: the relative risk aversion parameter is inversely related to the elasticity of intertemporal substitution: Epstein and Zin (1989) show how this restriction can be relaxed in a more general model with power utility where the timing of uncertainty resolution matters (see also Epstein and Zin 1991; Attanasio and Weber 1989).

Interestingly, an Euler equation for an asset holds even if there are important imperfections in some other assets. As long as the consumer is exploiting a given margin to move resources over time, an equation such as 7 will apply. If the interest rate for a given asset changes with the level of the asset, then the Euler Eq. (7) will have to be augmented with a term reflecting this effect (Pissarides 1978).

Liquidity Constraints

The Euler equation will be violated when the consumer is able, for some reason, to borrow against future income. In such a situation, Eq. (7) will hold as an inequality and the marginal utility of current consumption will be higher than the present discounted value of future consumption. Consumers who are liquidity constrained will be very sensitive to changes in current income. This case has received a considerable amount of attention in the literature. Many of the tests of violation of the Euler equation, such as Zeldes (1989), have focused on the so-called 'excess sensitivity' of consumption changes to predictable changes in income. It should be mentioned that, in a model with finite lives and a non-zero probability that income would be zero in each time period, standard regularity conditions on the utility function imply that a consumer will never want to borrow. If income is bounded away from zero, then the maximum the consumer will want to borrow is the present discounted value of the minimum value of income repeated in the future. This type of constraint has been sometimes referred to as a 'natural' liquidity constraint. Notice that such a constraint does not imply a violation of the Euler equation. If the restriction to borrowing is tighter, the Euler equation will instead be occasionally violated. And, even in periods in which it is not violated, the level of consumption will be affected by the possibility that the constraint will be binding in the future. As Hayashi (1987) explains, the presence of an operative, albeit not binding, liquidity constraint is equivalent to a shortening of the planning horizon or an increase in the discount rate. Evidence can be obtained by noting that consumers who are liquidity constrained will not be sensitive to changes in the level of the interest rate. As they will be at a kink of an intertemporal budget constraint, the demand for loans will be inelastic to changes in the slope of such an intertemporal budget constraint: the interest rate.

Endogenous Liquidity Constraint

In recent years, several studies have tried to model the shortcomings of credit and insurance markets by allowing for specific imperfections and frictions explicitly. The two main causes of imperfections that have been considered are: (*a*) private and asymmetric information and (b) the inability to perfectly enforce contracts. Models of this type can be seen as ways to endogenize specific market structures (such as one where consumers have access to a single asset in which they cannot borrow). In an influential paper, for instance, Cole and Kocherlakota (2001) show that an economy where individuals have a single bond in which they can borrow can be derived as a constrained equilibrium outcome where individuals have private information both on their income and on their savings.

Further Extensions and Alternative Models

While the evidence on the relevance of the life-cycle model is still inconclusive, a number of empirical puzzles have directed attention to more complex preference structures. In particular, the equity premium puzzle and the evolution of aggregate saving rates in high-growth economies (South East Asia) has led macroeconomists to incorporate habits into the model. However, there is still little formal evidence on the empirical relevance of habits in micro data. The widely documented retirement consumption puzzle (that is, a sudden drop of consumption at retirement) as well as a number of more or less anecdotal pieces of evidence on the inadequacy of saving for retirement and other forms of 'irrational' behaviour, have been interpreted as potentially supportive of time-inconsistent preferences. The most elegant way to introduce time-inconsistent preferences is provided by the hyperbolic discounting assumption (Laibson 1997).

Habits

Habits cause consumers to adjust slowly to shocks to permanent income, thus potentially explaining the excess smoothness of aggregate consumption, but also increase the utility loss associated with consumption drops, and may therefore help explain the equity premium puzzle.

Habits can take various forms: today's marginal utility may depend on the consumer's own past consumption level (internal habits) or the past consumption level of other consumers (external habits). This latter model seems to work better on aggregate data (Campbell and Cochrane 1999), even though a recent survey by Chen and Ludvigson (2004) challenges this conclusion.

Empirical macro-evidence on the presence of habits is mixed, and this may be due to the very nature of aggregate consumption data, as stressed in Dynan (2000). The serial correlation of aggregate consumption growth is affected by time aggregation (Heaton 1993), by aggregation over consumers, and by data construction methods (particularly for the services from durable goods). For this reason micro data seem preferable.

The simplest way to introduce habits (or durability) of consumption is to write the utility function as follows:

$$\Sigma_t u(x_t - \gamma' x_{t-1}; z_t) \tag{9}$$

where x is a vector of goods or services and z is any other variable that affects marginal utility (demographics, leisure, other goods that are not explicitly modelled). The γ parameters are positive for goods that provide services across periods (durability), negative for goods that are addictive (habit formation) or zero for goods that are fully non-durable, non-habit forming (Hayashi 1985).

The Euler equations corresponding to (9) involves x at four different periods of time, and their estimation typically requires panel data. High-quality consumption panel data are rare, and this has limited the scope for empirical analysis. Meghir and Weber (1996) have used Consumer Expenditure Survey (CEX) quarterly data on food, transport and services (and a more flexible specification of intertemporal non-separabilities than is implied by Eq. 9), and found no evidence of either durability or habits once leisure, stock of durables and cars as well as other conditioning variables are taken into consideration.

Similarly negative evidence on habits has been reported by Dynan (2000), using Panel Study of Income Dynamics (PSID) annual food at home data. Carrasco et al. (2005) use Spanish panel data and find some evidence for habits.

The few studies that have used micro data on non-durable consumption items to investigate the issue find little or no evidence of habits, at least once preferences capture the presence of non-separabilities between goods and leisure.

Durable Goods

The presence of durable goods has received less attention in the micro-based literature than in the macro-literature, which has stressed the importance of their high volatility to explain business cycle fluctuations (Mankiw 1982; Chah et al. 1995).

The simplest way to introduce durable goods into the analysis is to let the stock of durables affect utility (on the assumption that services are proportional to the stock), and to posit a relation between current stock, S_t, previous stock, S_{t-1}, and current purchases q_t (or maintenance and repairs) in physical terms like:

$$S_t = (1 - \rho)S_{t-1} + q_t \tag{10}$$

where ρ is a constant depreciation rate. This leads to the standard first-order condition for the durable good, according to which the relevant price is the user cost.

Typically, durable goods are costly to adjust, because of transaction costs (resale markets are dominated by information problems, known as the 'lemon' problem, and search costs are non-negligible). Sometimes these costs are modelled as a convex, differentiable function (Bernanke 1985), but the recent literature has stressed the need to take into account their non-differentiable nature (Grossman and Laroque 1990; Eberly 1994; Attanasio 2000; Bertola et al. 2005). This generates infrequent adjustment: consumers do not adjust continuously in response to depreciation, or income and price shocks, but wait until the actual stock hits either a lower limit, s, or an upper limit, S, and then adjust it to a target level. An interesting feature of this literature is that aggregate behaviour reflects changes in both the number of consumers that adjust and in the target level.

Durable goods might also play an insurance role, because they can be used to sustain consumption when times are bad. Postponing the purchase of food, or clothing, is certainly harder than failing to replace an old refrigerator or car, and housing maintenance can be put off for very long periods before structural damage occurs (Browning and Crossley 2000). Durable goods also play a more specific insurance role, against changes in the price of the corresponding services. This is particularly relevant in the case of housing, where owning your home may be the best way to hedge the risk of future increases in the market price of housing services (Sinai and Souleles 2005). Durable goods can also play a liquidity role, if they can be used as collateral to obtain a loan that pays for current consumption (Alessie et al. 1997). A typical example could be the ability to remortgage a house, or to borrow 100 per cent of the value of a newly purchased car.

Even if one is not interested in modelling durable goods, the existence of a stock of durables should not be neglected when estimating preference parameters if utility is not additive in non-durable goods and durable services. Significant effects of durable goods (cars) on the Euler equation for non-durables have been found in UK data (Alessie et al. 1997), and US data (Padula 1999).

Quasi-Hyperbolic Discounting

The widely documented consumption puzzle (that is the sudden drop of consumption at retirement, see Hamermesh 1984; Banks et al. 1998; and Bernheim et al. 2001), as well as a number of more or less anecdotal pieces of evidence on the inadequacy of saving for retirement and other forms of 'irrational' behaviour, have been interpreted as potentially supportive of time-inconsistent preferences. The most elegant way to introduce time-inconsistent preferences is provided by the quasi-hyperbolic discounting assumption (Laibson 1997). Consumers maximize the expected value of the following life-time utility index:

$$u(c_t) + \beta \sum_{\tau=1}^{T-t} \delta^t u(c_{t+\tau}) \tag{11}$$

This implies that a different, lower discount factor is used to choose between this period and the next (the product of β and δ) and between any two other periods (δ). This generates time-inconsistent plans, with too little saving for retirement. For this reason, consumers may choose to enter long-term commitment plans, such as 401(k) s in the United States.

The quasi-hyperbolic discounting model lends itself to estimation and testing, but requires solving for the consumption function numerically. Even though an Euler equation for this model has been derived, its empirical use is limited, because it involves the marginal propensity to consume out of wealth (Harris and Laibson

2001). It also suffers from some potential difficulties related to the definition of the time period, which crucially affects the properties of the solution, the length of which is arbitrarily set by the researcher.

A more tractable specification of preferences that may be used to model quasi-rational impatience has been put forward by Gul and Pesendorfer (2001, 2004), who stress the importance of self-control problems leading to the postponement of saving.

Where Do we Stand?

Since the 1970s we have learned much about the empirical implications of the life-cycle model and about the details of the model that need to be modified to fit the available evidence. Much work, however, remains to be done. In particular, there is scope to develop more complex numerical models that incorporate several realistic features. The areas of labour supply and housing are, in our opinion, particularly important. We also need to develop our understanding of the empirical implications of alternative models, such as hyperbolic discounting and check the extent to which they are empirically distinguishable from more standard models with complex preferences. Finally, it is important to stress the need for more and better data. One of the lessons learned from the development of new surveys that have been used to measure household wealth is that with enough ingenuity and creativity one can measure several of the variables that are relevant for our understanding of consumption and saving behaviour.

Our analysis of consumption and saving requires that more comprehensive measures of consumption are included in existing surveys, and that we learn to make systematic use of records on expectations, perceived uncertainty and so on.

See Also

- ▶ Consumer Expenditure
- ▶ Consumption-Based Asset Pricing Models (Empirical Performance)
- ▶ Consumption-Based Asset Pricing Models (Theory)
- ▶ Elasticity of Intertemporal Substitution
- ▶ Engel Curve
- ▶ Friedman, Milton (1912–2006)
- ▶ Modigliani, Franco (1918–2003)
- ▶ Precautionary Saving and Precautionary Wealth
- ▶ Revealed Preference Theory

Bibliography

Abel, A.B.. 1990. Asset prices under habit formation and catching up with the Joneses. *American Economic Review* 80: 38–42.

Alessie, R., M.P. Devereux, and G. Weber. 1997. Intertemporal consumption, durables and liquidity constraints: A cohort analysis. *European Economic Review* 41: 37–59.

Atkenson, A., and M. Ogaki. 1996. Wealth varying intertemporal elasticities of substitution: Evidence from panel and aggregate data. *Journal of Monetary Economics* 38: 507–534.

Attanasio, O.P. 2000. Consumer durables and inertial behaviour: Estimation and aggregation of (S, s) rules for automobile purchases. *Review of Economic Studies* 67: 667–696.

Attanasio, O.P., and M. Browning. 1995. Consumption over the life cycle and over the business cycle. *American Economic Review* 85: 1118–1137.

Attanasio, O.P., and A. Brugiavini. 2003. Social security and households' saving. *Quarterly Journal of Economics* 118: 1075–1119.

Attanasio, O.P., and S.J. Davis. 1996. Relative wage movements and the distribution of consumption. *Journal of Political Economy* 104: 1227–1262.

Attanasio, O.P., and H. Low. 2004. Estimating Euler equations. *Review of Economic Dynamics* 7: 405–435.

Attanasio, O.P., and S. Rohwedder. 2003. Pension wealth and household saving: Evidence from pension reforms in the United Kingdom. *American Economic Review* 93: 1499–1521.

Attanasio, O.P., and G. Weber. 1989. Intertemporal substitution, risk aversion and the Euler equation for consumption. *Economic Journal* 99: 59–73.

Attanasio, O.P., and G. Weber. 1993. Consumption growth, the interest rate and aggregation. *Review of Economic Studies* 60: 631–649.

Attanasio, O.P., and G. Weber. 1994. The UK consumption boom of the late 1980s: Aggregate implications of microeconomic evidence. *Economic Journal* 104: 1269–1302.

Attanasio, O.P., and G. Weber. 1995. Is consumption growth consistent with intertemporal optimization?

Evidence from the consumer expenditure survey. *Journal of Political Economy* 103: 1121–1157.

Attanasio, O.P., J. Banks, C. Meghir, and G. Weber. 1999. Humps and bumps in life-time consumption. *Journal of Business and Economic Statistics* 17: 22–35.

Banks, J., R. Blundell, and S. Tanner. 1998. Is there a retirement-savings puzzle? *American Economic Review* 88: 769–788.

Bernanke, B. 1985. Adjustment costs, durables and aggregate consumption. *Journal of Monetary Economics* 15: 41–68.

Bernheim, B.D., J. Skinner, and S. Weinberg. 2001. What accounts for the variation in retirement wealth among U.S. households? *American Economic Review* 91: 832–857.

Bertola, G., L. Guiso, and L. Pistaferri. 2005. Uncertainty and consumer durables adjustment. *Review of Economic Studies* 72: 973–1007.

Bewley, T.F. 1977. The permanent income hypothesis: A theoretical formulation. *Journal of Economic Theory* 16: 252–259.

Blinder, A., and A. Deaton. 1985. The time series consumption function revisited. *Brookings Papers on Economic Activity* 1985(2): 465–521.

Blundell, R., M. Browning, and C. Meghir. 1994. Consumer demand and the life-cycle allocation of household expenditures. *Review of Economic Studies* 61: 57–80.

Bodie, Z., R.C. Merton, and W.F. Samuelson. 1992. Labor supply flexibility and portfolio choice in a life-cycle model. *Journal of Economic Dynamics and Control* 16: 427–449.

Boskin, M.J. 1978. Taxation, saving and the rate of interest. *Journal of Political Economy*: S3–S28.

Breeden, D.T. 1979. An intertemporal asset pricing model with stochastic consumption and investment opportunities. *Journal of Financial Economics* 7: 265–296.

Browning, M., and T.F. Crossley. 2000. Luxuries are easier to postpone: A proof. *Journal of Political Economy* 108: 1022–1026.

Browning, M., and M. Ejrnaes. 2002. Consumption and children. Working Paper No. 2002–6, Centre for Applied Microeconometrics, University of Copenhagen.

Caballero, R.J. 1990. Consumption puzzles and precautionary savings. *Journal of Monetary Economics* 25: 113–136.

Caballero, R.J. 1991. Earnings uncertainty and aggregate wealth accumulation. *American Economic Review* 81: 859–871.

Campbell, J.Y. 1987. Does saving anticipate declining labor income? An alternative test of the permanent income hypothesis. *Econometrica* 55: 1249–1273.

Campbell, J.Y., and J. Cochrane. 1999. Force of habit: A consumption-based explanation of aggregate stock market behavior. *Journal of Political Economy* 107: 205–251.

Campbell, J.Y., and A. Deaton. 1989. Why is consumption so smooth? *Review of Economic Studies* 56: 357–373.

Campbell, J.Y., and N.G. Mankiw. 1990. Permanent income, current income, and consumption. *Journal of Business and Economic Statistics* 8: 265–279.

Campbell, J.Y., and N.G. Mankiw. 1991. The response of consumption to income: A cross-country investigation. *European Economic Review* 35: 723–756.

Carrasco, R., J.M. Labeaga, and J.D. López-Salido. 2005. Consumption and habits: Evidence from panel data. *Economic Journal* 115: 144–165.

Carroll, C.D. 1992. The buffer-stock theory of saving: Some macroeconomic evidence. *Brookings Papers on Economic Activity* 1992(2): 61–156.

Carroll, C.D. 1997. Buffer-stock saving and the life cycle/permanent income hypothesis. *Quarterly Journal of Economics* 112: 1–55.

Carroll, C.D., and M.S. Kimball. 1996. On the concavity of the consumption function. *Econometrica* 64: 981–992.

Chah, E.Y., V.A. Ramey, and R.M. Starr. 1995. Liquidity constraints and intertemporal consumer optimization: Theory and evidence from durable goods. *Journal of Money, Credit and Banking* 27: 272–287.

Chen, X., and S.C. Ludvigson. 2004. *A land of addicts? An empirical investigation of habits-based asset pricing models*. New York: New York University Press.

Choi, J.J., D. Laibson, B.C. Madrian, and A. Metrick. 2006. Saving for retirement on the path of least resistence. In *Behavioral public finance: Toward a New Agenda*, ed. E. McCaffrey and J. Slemrod. New York: Russell Sage Foundation.

Cochrane, J.H. 1991. A simple test of consumption insurance. *Journal of Political Economy* 99: 957–976.

Cole, H.L., and N.R. Kocherlakota. 2001. Efficient allocations with hidden income and hidden storage. *Review of Economic Studies* 68: 523–542.

Deaton, A. 1991. Saving and liquidity constraints. *Econometrica* 59: 1221–1248.

Dynan, K. 2000. Habit formation in consumer preferences: Evidence from panel data. *American Economic Review* 90: 391–406.

Eberly, J.C. 1994. Adjustment of consumers' durables stocks: Evidence from automobile purchases. *Journal of Political Economy* 102: 403–436.

Epstein, L.G., and S.E. Zin. 1989. Substitution, risk aversion, and the temporal behavior of consumption and asset returns: A theoretical framework. *Econometrica* 57: 937–969.

Epstein, L.G., and S.E. Zin. 1991. Substitution, risk aversion, and the temporal behavior of consumption and asset returns: An empirical analysis. *Journal of Political Economy* 99: 263–286.

Flavin, M.A. 1981. The adjustment of consumption to changing expectations about future income. *Journal of Political Economy* 89: 974–1009.

Friedman, M. 1957. *A theory of the consumption. function*. Princeton: Princeton University Press.

Gollier, C. 1995. The comparative statics of changes in risk revisited. *Journal of Economic Theory* 66: 522–536.

Gourinchas, P.-O., and J.A. Parker. 2002. Consumption over the life cycle. *Econometrica* 70: 47–89.

Grossman, S.J., and G. Laroque. 1990. Asset pricing and optimal portfolio choice in the presence of illiquid durable consumption goods. *Econometrica* 58: 25–51.

Gul, F., and W. Pesendorfer. 2001. Temptation and self-control. *Econometrica* 9: 1403–1435.

Gul, F., and W. Pesendorfer. 2004. Self-control and the theory of consumption. *Econometrica* 72: 119–158.

Hall, R.E. 1978. Stochastic implications of the life cycle-permanent income hypothesis: Theory and evidence. *Journal of Political Economy* 86: 971–987.

Hall, R.E. 1988. Intertemporal substitution in consumption. *Journal of Political Economy* 96: 339–357.

Hall, R.E., and F.S. Mishkin. 1982. The sensitivity of consumption to transitory income: Estimates from panel data on households. *Econometrica* 50: 461–481.

Hamermesh, D.S. 1984. Consumption during retirement: The missing link in the life cycle. *Review of Economics and Statistics* 66: 1–7.

Hansen, L.P., and K.J. Singleton. 1982. Generalized instrumental variables estimation of nonlinear rational expectations models. *Econometrica* 50: 1269–1286.

Hansen, L.P., and K.J. Singleton. 1983. Stochastic consumption, risk aversion, and the temporal behavior of asset returns. *Journal of Political Economy* 91: 249–265.

Harris, C., and D. Laibson. 2001. Dynamic choices of hyperbolic consumers. *Econometrica* 69: 935–957.

Hayashi, F. 1985. The permanent income hypothesis and consumption durability: Analysis based on Japanese panel data. *Quarterly Journal of Economics* 100: 1083–1113.

Hayashi, F. 1987. Tests for liquidity constraints: A critical survey. In *Advances in econometrics II: Fifth world congress*, ed. T. Bewley. Cambridge: Cambridge University Press.

Hayashi, F., J. Altonji, and L. Kotlikoff. 1996. Risk-sharing between and within Families. *Econometrica* 64: 261–294.

Heaton, J. 1993. The interaction between time-nonseparable preferences and time aggregation. *Econometrica* 61: 353–385.

Heckman, J.J. 1974. Life cycle consumption and labor supply: An explanation of the relationship between income and consumption over the life cycle. *American Economic Review* 64: 188–194.

Hubbard, R.G., J. Skinner, and S.P. Zeldes. 1994. The importance of precautionary motives in explaining individual and aggregate saving. *Carnegie-Rochester Conference Series on Public Policy* 40: 59–125.

Hubbard, R.G., J. Skinner, and S.P. Zeldes. 1995. Precautionary saving and social insurance. *Journal of Political Economy* 103: 360–399.

King, R.G., and C.I. Plosser. 1984. Money, credit, and prices in a real business cycle. *American Economic Review* 74: 363–380.

Laibson, D. 1997. Golden eggs and hyperbolic discounting. *Quarterly Journal of Economics* 62: 443–477.

Lucas, R.E. Jr. 1978. Asset prices in an exchange economy. *Econometrica* 46: 1429–1445.

MaCurdy, T.E. 1981. An empirical model of labor supply in a life-cycle setting. *Journal of Political Economy* 89: 1345–1370.

MaCurdy, T.E. 1999. An essay on the life cycle: Characterizing intertemporal behavior with uncertainty, taxes, human capital, durables, imperfect capital markets, and nonseparable preferences. *Research in Economics* 53: 5–46.

Mankiw, G.N. 1982. Hall's consumption hypothesis and durable goods. *Journal of Monetary Economics* 10: 417–425.

Meghir, C., and G. Weber. 1996. Intertemporal nonseparability or borrowing restrictions? A disaggregate analysis using a U.S. consumption panel. *Econometrica* 64: 1151–1181.

Miniaci, R., and G. Weber. 1999. The Italian recession of 1993: Aggregate implications of microeconomic evidence. *Review of Economics and Statistics* 81: 237–249.

Modigliani, F., and R. Brumberg. 1954. Utility analysis and the consumption function: An interpretation of cross-section data. In *Post Keynesian economics*, ed. K. Kurihara. New Brunswick: Rutgers University Press.

Padula, M. 1999. Euler equations and durable goods. Working Paper No. 30, CSEF.

Parker, J.A. 1999. The reaction of household consumption to predictable changes in social security taxes. *American Economic Review* 89: 959–973.

Pissarides, C.A. 1978. Liquidity considerations in the theory of consumption. *Quarterly Journal of Economics* 92: 279–296.

Poterba, J. 1988. Are consumers forward looking? Evidence from fiscal experiments. *American Economic Review* 78: 413–418.

Sargent, T.J. 1978. Rational expectations, econometric exogeneity, and consumption. *Journal of Political Economy* 86: 673–700.

Shapiro, M.D., and J. Slemrod. 2003. Consumer response to tax rebates. *American Economic Review* 93: 381–396.

Sinai, T., and N.S. Souleles. 2005. Owner-occupied housing as a hedge against rent risk. *Quarterly Journal of Economics* 120: 763–789.

Souleles, N.S. 1999. The response of household consumption to income tax refunds. *American Economic Review* 89: 947–958.

Summers, L. 1981. Capital taxation and capital accumulation in a life-cycle growth model. *American Economic Review* 71: 533–544.

Thurow, L.C. 1969. The optimum lifetime distribution of consumption expenditures. *American Economic Review* 59: 324–330.

Townsend, R.M. 1994. Risk and insurance in village India. *Econometrica* 62: 539–591.

West, K.D. 1988. The insensitivity of consumption to news about income. *Journal of Monetary Economics* 21: 17–33.

Zeldes, S.P. 1989. Consumption and liquidity constraints: An empirical investigation. *Journal of Political Economy* 97: 305–346.

C

Consumer Surplus

Daniel T. Slesnick

Abstract

Over the years, consumer surplus has been used to measure the welfare effects of price and income changes. Despite its widespread use, it provides a measure of well-being that is ordinally equivalent to the change in utility only under conditions that are inconsistent with long-standing empirical evidence. Hicksian surplus measures, such as the equivalent or compensating variations, provide exact indicators of the change in utility without such restrictions. Beginning in the early 1980s, empirical methods have been developed to estimate the equivalent variation that has the same data requirements as consumer surplus.

Keywords

Aggregation; Compensating variation; Consumer surplus; Equivalent variation; Expenditure function; Indirect utility function; Integrability of demand; Intertemporal welfare effects; Linear expenditure system; Marginal utility of income; Representative agent; Roy's identity; Social choice; Social expenditure function; Well-being

JEL Classifications

D11

How does the market power exercised by firms influence consumer welfare? What is the effect of excise taxes on households with different levels of income? Does governmental regulation increase the welfare of consumers? Topical issues such as these indicate that the measurement of welfare is a fundamental element of public policy analysis. Indeed, a full consideration of taxes, subsidies, transfer programmes, health care reform, regulation, environmental policy, the social security system, and educational reform must ultimately address the question of how these policies affect individual well-being.

While centrally important to many problems of economic analysis, confusion persists concerning the relationship between commonly used indicators of welfare and well-established theoretical formulations. For more than 150 years, consumer surplus has been used to measure the welfare effects of changes in prices and incomes. Its popularity can be ascribed to its intuitive appeal, the ease with which it is implemented, and its modest data requirements. Although it is generally accepted that Dupuit (1844) was the originator of the concept of consumer surplus, it is largely attributed to Marshall (1890). (Chipman and Moore (1976), provide a brief survey of the history of the debate related to consumer surplus.) We begin with the following notation:

$\mathbf{p} = (p_1, p_2, \ldots, p_n)$ – a vector of commodity prices.
Y_k – the income of individual k.
A_k – a vector of demographic characteristics of individual k.
$x_{ik} = x_i(\mathbf{p}, Y_k, \mathbf{A_k})$ is the demand for good i by individual k.

Suppose we are interested in the welfare impact of a change in the price of a single commodity from p_1^0 to p_1^1. The change in consumer surplus is given by:

$$\Delta CS_k = -\int_{p_1^0}^{p_1^1} x_1(t, p_2, \ldots p_n, Y_k, \mathbf{A_k})dt. \quad (1)$$

If ΔCS_k is positive (negative), the price change is judged to have increased (decreased) the welfare of individual k. Is it ordinally equivalent to the change in utility? A necessary condition is that the demand function is generated by a rational consumer who maximizes utility subject to a budget constraint. Unless consumers have optimized and are at the boundaries of their budget sets, it is impossible to assess the welfare effects of changes in prices and incomes. (That is, demands must be 'integrable' and consistent with a well-behaved utility function. Hurwicz and Uzawa (1971),

provide a formal statement of the integrability conditions.)

If demands are consistent with rational consumer behaviour, an indirect utility function $V(\mathbf{p}, Y_k, \mathbf{A_k})$ represents the maximum utility attained at prices $\mathbf{p}$ and income Y_k, and Roy's Identity provides the link between demands and utility:

$$x_1(\mathbf{p}, Y_k, \mathbf{A_k}) = -\frac{\partial V(\mathbf{p}, Y_k, \mathbf{A_k})/\partial p_1}{\partial V(\mathbf{p}, Y_k, \mathbf{A_k})/\partial Y_k}. \quad (2)$$

If the marginal utility of income is constant, substitution of (2) into (1) yields an explicit expression for the change in consumer surplus that is ordinally equivalent to the change in utility:

$$\Delta CS_k = \frac{V(\mathbf{p}^1, Y_k, \mathbf{A_k}) - V(\mathbf{p}^0, Y_k, \mathbf{A_k})}{\partial V/\partial Y_k}.$$

While constancy of the marginal utility of income is restrictive, Chipman and Moore (1976, 1980) have shown that application of consumer surplus is more problematical if there are changes in more than one price. In such circumstances, the change in consumer surplus must be evaluated using a line integral defined over the path of price changes from $\mathbf{p^0}$ to $\mathbf{p}^1$:

$$\Delta CS_k = \int_{p^0}^{p^1} \sum_i \mathrm{X}_i(\mathbf{p}, Y_k, \mathbf{A_k}) \mathbf{dp_i}. \quad (3)$$

Price paths are not observed so it is essential that (3) be path independent. This holds if the uncompensated price effects are symmetric (see, for example, Angus Taylor and Robert Mann 1972, pp. 500–4):

$$\frac{\partial x_i}{\partial p_j} = \frac{\partial x_j}{\partial p_i} \text{ for } all\ i \neq j.$$

This form of symmetry requires preferences to be homothetic, which is a restriction that is inconsistent with well-established empirical regularities.

In the most general circumstance of changes in prices and income, consumer surplus is defined as:

$$\Delta CS_k = -\int_Z \sum x_{\mathrm{i}}(\mathbf{p}, Y_k, \mathbf{A_k}) \mathrm{d}p_i + \left(Y_k^1 - Y_k^0\right), \quad (4)$$

where $\mathbf{Z}$ is a path between $\left(\mathbf{p}^0, Y_k^0\right)$ and $\left(\mathbf{p}^1, Y_k^1\right)$. Chipman and Moore (1976) have demonstrated that here are no circumstances under which (4) is path independent and ordinally equivalent to the change in utility of a rational consumer.

Hicksian Surplus Measures

Given the problems with consumer surplus, how should the welfare effects of price and income changes be measured? Hicks (1942) developed an approach that is exactly analogous to (4) once we substitute compensated for uncompensated demand functions:

$$\Delta HS_k = -\int_z \sum x_i^c(\mathbf{p}, V, \mathbf{A_k}) \mathrm{d}p_i + \left(Y_k^1 - Y_k^0\right) \quad (5)$$

where $x_i^c(\mathbf{p}, V, \mathbf{A_k})$ is the compensated demand for the ith good evaluated at utility level V. Compensated price effects are symmetric, so the line integral in (5) is path independent and the surplus measure is single-valued.

For simple binary comparisons of policies, the utility level at which ΔHS_k is evaluated is often treated as a matter of little consequence. If it is calculated at the utility attained at prices $\mathbf{p}^1$ and income Y_k^1 (denoted V^1), a generalized version of the equivalent variation is obtained:

$$\begin{aligned} EV_k &= E\left(\mathbf{p}^0, V^1, \mathbf{A_k}\right) \\ &\quad - E\left(\mathbf{p}^1, V^1, \mathbf{A_k}\right) + \left(Y_k^1 - Y_k^0\right) \\ &= E\left(\mathbf{p}^0, V^1, \mathbf{A_k}\right) - E\left(\mathbf{p}^0, V^0, \mathbf{A_k}\right) \end{aligned} \quad (6)$$

where $E(\mathbf{p}, V, \mathbf{A_k})$ is the expenditure function, defined as the minimum income needed for individual k to attain utility V at prices $\mathbf{p}$. Not only is the generalized equivalent variation single-valued, but it is ordinally equivalent to the change

in utility. That is, EV_k is positive if and only if $V^1 > V^0$.

The utility level at which (5) is evaluated is important for multiple comparisons of price and income changes. The generalized equivalent variation will give an ordering of outcomes that is identical to that based on utility levels. If (5) is evaluated at $V^0 = V(\mathbf{p}^0, Y_k^0, \mathbf{A_k})$, we obtain the generalized compensating variation:

$$\begin{aligned} CV_k &= E(\mathbf{p}^0, V^0, \mathbf{A_k}) - E(\mathbf{p}^1, V^0, \mathbf{A_k}) + (Y_k^1 - Y_k^0) \\ &= E(\mathbf{p}^1, V^1, \mathbf{A_k}) - E(\mathbf{p}^1, V^0, \mathbf{A_k}). \end{aligned} \tag{7}$$

Because the utility levels are 'cardinalized' using different prices for each set of binary comparisons, the ordering of multiple outcomes based on (7) need not match the ordering based on utility levels. Chipman and Moore (1980) have shown that consistent rankings of outcomes require restrictions on preferences that are the same as for consumer surplus.

While the simple static formulation of consumer surplus is the most frequent application, the conceptual framework can be extended to analyse the effects of changes in utility in more general settings. For example, intertemporal welfare effects are often represented as the discounted sum of the within-period equivalent or compensating variations.

Keen (1990) has shown that this will differ from the lifetime equivalent variation to the extent that individuals are able to substitute intertemporally. As an alternative approach, he defines V_L to be the maximum level of lifetime utility of an individual who lives T periods when the profiles of prices and interest rates are $\{\mathbf{p_t}\}$ and $\{r_t\}$ respectively. If the (optimal) time path of utility corresponding to V_L at these prices and interest rates is $\{V_{kt}\}$, the lifetime expenditure function can be represented as:

$$\Omega_L(\{\mathbf{p_t}\}, \{r_t\}, V_L) = \sum_t g_t \; E(\mathbf{p_t}, V_{kt}, \mathbf{A_{kt}}),$$

where $g_t = \Pi_{s=0}^{t}(1 + r_s)^{-1}$.

As in the static framework, the lifetime expenditure function can be used to represent an exact measure of the change in lifetime welfare. Define V_L^1 to be the maximum level of lifetime welfare when the profile of prices and interest rates are $\{\mathbf{p_t^1}\}$ and $\{r_t^1\}$ and denote the corresponding time path of utility as $\{V_{kt}^1\}$. The reference prices and interest rates, $\{\mathbf{p_t^0}\}$ and $\{r_t^0\}$, yield a lifetime utility level of V_L^0 and within-period utilities $\{V_{kt}^0\}$. Keen's exact measure of the change in lifetime welfare, evaluated at the reference prices, is exactly analogous to the generalized equivalent variation:

$$\begin{aligned} \Delta W_L = \Omega_L(\{\mathbf{p_t^0}\}, \{\mathrm{r}_t^0\}, V_L^1) \\ - \Omega_L(\{\mathbf{p_t^0}\}, \{r_t^0\}, V_L^0). \end{aligned}$$

The concepts of the equivalent and compensating variation can also be extended to cases in which the choices made by consumers are discrete rather than continuous. Dagsvik and Karlstrom (2005) describe the compensating variation in the context of a random utility model defined as:

$$U_{jk} = V(\mathbf{p_j}, \mathrm{Y}_k, \mathbf{A_k}) + \varepsilon_{\mathrm{jk}}(\mathrm{j} = 1, 2, \ldots, \mathrm{J}),$$

where U_{jk} is the utility of individual k in alternative j, $V(.)$ is a deterministic indirect utility function, and ε_{jk} are random variables. There are a total of J choices available to the consumer and, for simplicity, it is assumed that only prices vary across alternatives.

Consider the welfare effect of a change in the set of prices and income facing individual k from $(\mathbf{p}_1^0, \mathbf{p}_2^0, \ldots, \mathbf{p}_\mathbf{J}^0, Y_k^0)$ to $(\mathbf{p}_1^1, \mathbf{p}_2^1, \ldots, \mathrm{p}_\mathbf{J}^1, Y_k^1)$. If the consumer chooses the alternative that maximizes U_{jk}, the compensating variation is defined implicitly as that value CV_k that satisfies the following equality:

$$\begin{aligned} &\max_j V(\mathbf{p_j^0}, Y_\mathrm{k}^0, \mathbf{A_k}) + \varepsilon_{jk} \\ &\quad = \max_j V(\mathbf{p_j^1}, Y_k^1 - CV_k, \mathbf{A_k}) + \varepsilon_{jk}. \end{aligned}$$

Although conceptually analogous to the equivalent and compensating variation described previously, CV_k is now random and cannot, in general, be represented in closed form.

From Demand Functions to Welfare Measurement

While it was understood that the equivalent variation resolved the conceptual problem of welfare measurement, it had little influence on applied welfare economics because compensated demand functions were presumed to be unobservable. Willig (1976) made the first attempt to bridge the gap between theory and application by showing that, for a single price change, consumer surplus can provide an approximation to the equivalent or compensating variation. However, with multiple price and income changes, consumer surplus is not single-valued and is of no use in approximating changes in economic welfare (McKenzie 1979).

Shortly after the publication of Willig's paper, however, empirical procedures were developed to estimate the equivalent or compensating variation. Each method begins with the specification of a demand function and, under the assumption of integrability, is used to recover the utility or expenditure functions. The complexity of this procedure diminishes if demand functions are linear, and consideration is restricted to changes in the price of a single good.

Hausman (1981) provided an analytic solution to this problem for a demand function given by:

$$x_1 = \gamma_p p_1 + \gamma_Y Y_k + \gamma_{\mathbf{A}} \mathbf{A_k},$$

where γ_p, γ_Y, and $\gamma_{\mathbf{A}}$ are unknown parameters to be estimated econometrically. Roy's Identity provides a partial differential equation that can be solved to obtain an expenditure function of the form:

$$E(p_1, V, \mathbf{A_k}) = \mathrm{V}e^{\gamma_Y p_1} - (1/\gamma_Y)\left[\gamma_p p_1 + \left(\gamma_p/\gamma_Y\right) + \gamma_{\mathbf{A}} \mathbf{A_k}\right]. \quad (8)$$

The expenditure function allows the equivalent variation to be computed exactly as in (6) and Willig-type approximations are unnecessary. Hausman's method has the same data requirements as consumer surplus, and only linear regression methods are needed to estimate the unknown parameters.

Closed form solutions to the partial differential equation implied by Roy's Identity can be obtained for only a limited class of demand functions. An alternative approach is to begin with an assumed form of the indirect utility function and use Roy's Identity to obtain a system of demand equations. Since the form of the utility function is assumed from the outset, it is unnecessary to solve a complex system of partial differential equations.

Muellbauer (1974) provided an early example of this approach. He assumed that demands were consistent with a Stone-Geary utility function given by:

$$V(\mathbf{p}, Y_k) = \frac{(Y_k - \sum p_i \delta_i)}{\Pi p_i^{\alpha_i}} \quad (9)$$

where $\delta = (\delta_1, \delta_2, \ldots, \delta_n)$ and $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ are unknown parameters. The corresponding expenditure function is:

$$E(\mathbf{p}, V) = \sum p_i \delta_i + V\left(\Pi p_i^{\alpha_i}\right).$$

The unknown parameters can be estimated by fitting the linear expenditure system to household budget data:

$$p_i x_i = p_i \delta_i + \alpha_i \left(Y_k - \sum p_i \delta_i\right) (i = 1, 2, \ldots, n). \quad (10)$$

Given estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}$, the expenditure function can be used to compute the equivalent or compensating variation as in (6) and (7).

While this is more general than Hausman's approach, it has its own disadvantages. For an assumed form of the utility function, the functional forms of the demands are the same for every good, which may hinder the ability of the model to fit the data. Is it possible to start with an arbitrary demand system (rather than a utility function) and measure the welfare effects of multiple price changes? Two elegant procedures were proposed that required more complicated calculations to recover the expenditure function, but did not impose restrictions on the form of the demand

functions other than the standard integrability conditions.

The first method is based on an approximation to McKenzie's (1957) indirect money metric utility function defined as:

$$\mu\left(\mathbf{p}, Y_k, \mathbf{A_k}; \mathbf{p}^0\right) = E\left(\mathbf{p}^0, V(\mathbf{p}, Y_k, \mathbf{A_k}), \mathbf{A_k}\right).$$

McKenzie and Pearce (1976) showed that $\Delta\mu$ can be approximated by a Taylor's series expansion about the initial equilibrium:

$$\begin{aligned}\Delta\mu = &\frac{\partial\mu}{\partial\mathbf{p}'}\Delta\mathbf{p}(1/2) + \Delta\mathbf{p}'\frac{\partial^2\mu}{\partial p\partial p'}\Delta\mathbf{p} \\ &+ \left(\frac{\partial\mu}{\partial Y} + \frac{\partial^2\mu}{\partial\mathbf{p}\partial Y'}\Delta\mathbf{p} + 1/2\frac{\partial^2\mu}{\partial Y^2}\Delta Y\right)\Delta Y \\ &+ R\end{aligned} \tag{11}$$

where R represents higher order terms in the series.

The expression in (11) can be represented as a function of uncompensated demand functions when μ is evaluated at the reference prices (this follows from Roy's Identity and from the fact that at these prices the marginal utility of income is equal to one and all higher income derivatives are zero – see McKenzie and Pearce 1976, for details):

$$\begin{aligned}\Delta\mu = &-\mathbf{X}'\Delta\mathbf{p} \\ &- (1/2)\Delta\mathbf{p}'\left(\frac{\partial\mathbf{X}}{\partial\mathrm{p}} - \mathbf{X}\frac{\partial\mathbf{X}}{\partial\mathbf{Y}'}\right)\Delta\mathbf{p} \\ &+ \left(1 - \frac{\partial\mathbf{X}}{\partial\mathbf{Y}'}\Delta\mathbf{p}\right)\Delta Y + R.\end{aligned} \tag{12}$$

Given knowledge of the demand functions and the magnitudes of the price and income effects, one has all of the information necessary to get as accurate an estimate of the change in utility as desired.

Vartia (1983) developed an algorithm that recovers the expenditure function numerically to any desired level of accuracy. Let $\mathbf{p}(t)$ and $Y_k(t)$ be the paths of price and income changes for $0 \leq t \leq 1$. As prices and income change, the movements of demands along an indifference curve can be represented implicitly by the differential equation:

$$\frac{dY_k(t)}{dt} = \sum_{i=1}^{n} x_i(\mathbf{p}(t), Y_k(t), \mathbf{A_k})\frac{dp_i(t)}{dt}.$$

Integrating over t yields an expression that can, in principle, be solved to obtain $E(\mathbf{p}(t), V^0, \mathbf{A_k})$ which is the centrepiece of the welfare calculations:

$$\begin{aligned}&E\left(\mathbf{p}(t), V^0, \mathbf{A_k}\right) - E\left(\mathbf{p}^0, V^0, \mathbf{A_k}\right) \\ &\quad = \sum_{i=1}^{n}\int_0^t x_i\left(\mathbf{p}(t), E\left(\mathbf{p}(t), V^0, \mathbf{A_k},\right), \mathbf{A_k}\right)\frac{dp_i(t)}{dt}dt.\end{aligned} \tag{13}$$

Vartia described several algorithms that can be used to solve this equation numerically over the price path $\mathbf{p}(t)$ so that, when evaluated at $t = 1$, we obtain $E(\mathbf{p}^1, V^0, \mathbf{A_k})$. As long as the demands satisfy the integrability conditions, the solution to (13) will be independent of the price path used in the algorithm. This method is valid for multiple price and expenditure changes and, because a closed-form solution is unnecessary, facilitates flexibility in estimating demand patterns.

Aggregation

The methods described to this point provide estimates of the change in welfare for individuals. In practice, analysts are more concerned about the impact of policies on groups. Micro-level estimates are an essential first step, but, for welfare economics to be useful to practitioners, a method of aggregation is essential. The easiest approach is to assume that market demands are generated by a representative consumer. Under this condition, the methods described previously can be applied to aggregate demands and the utility function of the representative agent can be recovered.

While frequently applied, this is unsatisfactory for a number of reasons. Market demands need not be consistent with a rational representative consumer. Even if every individual has demands that are consistent with utility maximization, aggregate demands need not satisfy any of the integrability conditions other than homogeneity of degree zero in prices and income (Sonnenschein 1972). Moreover, it is unclear what this utility function actually represents. Kirman (1992) presents an example in which the representative agent prefers (aggregate) market basket A to B even though all individuals prefer the reverse. This violation of the most basic principle of social choice suggests that the utility of the representative agent should not be used for policy analysis even in the unlikely event that aggregate demands are integrable.

An alternative approach is to define aggregate welfare to be a function of the individual surplus measures. Such an approach was advocated by Harberger (1971) in his effort to make consumer surplus the standard tool for applied welfare analysis. At a conceptual level, such an indicator of aggregate welfare appears to be a natural extension of the positive analysis of welfare measurement at the micro level. This is obviously not the case because aggregation necessitates normative judgements in which the gains to some must be weighed against the losses to others. Simply summing the surplus measures, for example, embodies a version of utilitarianism and ignores distributional concerns.

Since any method of measuring welfare for groups of individuals necessarily involves subjective judgements, it seems reasonable to state explicitly the underlying ethical basis for the method of ordering outcomes in the aggregate. The social choice theoretic framework used by Sen (1970) provides a reasonable way of presenting the normative assumptions related to the measurability and comparability of individual welfare levels that facilitate well-behaved social orderings of outcomes. Under conditions described by Sen and others, these orderings can be represented by a social welfare function:

$$W = W(V_1, V_2, \ldots, V_k)$$

where V_k is a welfare indicator of individual k.

A monetary measure of social welfare can be obtained using Pollak's (1981) concept of a social expenditure function:

$$M(\mathbf{p}, W) = \min\left\{Y : W(V_1, \ldots, V_k) \geq W, \sum Y_k = Y\right\}.$$

This function is exactly analogous to its micro-level counterpart and is the minimum level of aggregate income required to attain a specified social welfare contour. If W^0 is the social welfare under policy 0 and W^1 is the welfare under policy 1, the monetary measure of the change is social welfare is exactly analogous to the generalized equivalent variation:

$$\Delta W = M\left(\mathbf{p}, W^1\right) - M\left(\mathbf{p}, W^0\right).$$

ΔW is clearly ordinally equivalent to the changes in social welfare, and normative judgements are represented explicitly through the specification of the social welfare function.

See Also

- Cost–Benefit Analysis
- Cost Minimization and Utility Maximization
- Hicksian and Marshallian Demands
- Indirect Utility Function
- Social Welfare Function

Bibliography

Chipman, J.S., and J. Moore. 1976. The scope of consumer's surplus arguments. In *Evolution, welfare and time in economics: Essays in honor of Nicholas Georgescu-Roegen*, ed. A.M. Tang, F.M. Westfield, and J.S. Worley. Lexington: Heath-Lexington Books.

Chipman, J.S., and J. Moore. 1980. Compensating variation, consumer's surplus, and welfare. *American Economic Review* 70: 933–949.

Dagsvik, J., and A. Karlstrom. 2005. Compensating variation and Hicksian choice probabilities in random utility models that are nonlinear in income. *Review of Economic Studies* 72: 57–76.

Dupuit, J. 1844. On the measurement of the utility of public works. Trans R.H. Barback. In *Readings in Welfare Economics,* ed. K.J. Arrow and T. Scitovsky. Homewood: Richard D. Irwin.
Harberger, A.C. 1971. Three basic postulates for applied welfare economics. *Journal of Economic Literature* 9: 785–797.
Hausman, J.A. 1981. Exact consumer's surplus and deadweight loss. *American Economic Review* 71: 662–676.
Hicks, J.R. 1942. Consumer's surplus and index numbers. *Review of Economic Studies* 9 (2): 126–137.
Hurwicz, L., and H. Uzawa. 1971. On the integrability of demand functions. In *Preferences, utility and demand*, ed. J. Chipman et al. New York: Harcourt, Brace and Jovanovich.
Keen, M. 1990. Welfare analysis and intertemporal substitution. *Journal of Public Economics* 42: 47–66.
Kirman, A.P. 1992. Whom or what does the representative individual represent? *Journal of Economic Perspectives* 6 (2): 117–136.
Marshall, A. 1890. *Principles of economics*. London: Macmillan.
McKenzie, L. 1957. Demand theory without a utility index. *Review of Economic Studies* 24 (65): 185–189.
McKenzie, G.W. 1979. Consumer's surplus without apology: Comment. *American Economic Review* 69: 465–468.
McKenzie, G.W., and I. Pearce. 1976. Exact measures of welfare and the cost of living. *Review of Economic Studies* 43: 465–468.
Muellbauer, J. 1974. Prices and inequality: The U.K. experience. *Economic Journal* 84 (333): 32–55.
Pollak, R.A. 1981. The social cost of living index. *Journal of Public Economics* 15: 311–336.
Sen, A.K. 1970. *Collective choice and social welfare*. San Francisco: Holden Day.
Sonnenschein, H. 1972. Market excess demand functions. *Econometrica* 40: 549–563.
Taylor, A.E., and R. Mann. 1972. *Advanced calculus*. 2nd ed. Lexington: Xerox College Publishing.
Vartia, Y.O. 1983. Efficient methods of measuring welfare change and compensated income in terms of ordinary demand functions. *Econometrica* 51: 79–98.
Willig, R.E. 1976. Consumer's surplus without apology. *American Economic Review* 66: 589–597.

Consumption and Production

C. A. Gregory

Neoclassical economic analysis is carried out within a conceptual framework that views the economic process as a 'one way avenue' leading from 'factors of production' to 'consumption goods' (Sraffa 1960, p. 93). This stands in striking contrast to the approach of classical political economy which views the system of consumption and production as a circular process. This perspective was first developed by Quesnay (1759) and elaborated by Marx (1859) in his analysis of the economy in general. Marx developed a distinction between 'production and productive consumption' and 'consumption and consumptive production' and related this to the concepts of exchange and distribution. This distinction fell into diuse with the rise of neoclassical economics but has been rehabilitated by Sraffa (1960) in his famous critique of modern economics. The concept 'production and productive consumption' provides the general conceptual framework within which his particular theory of commodities is elaborated (see Sraffa 1960, p. 3). Sraffa's exposition not only advances our understanding of the theory of commodities, it also enables us to grasp the essence of Marx's important distinction between consumption and production. Marx expressed himself in rather obscure Hegelian terms and Sraffa's simple numerical examples clarify much of Marx's argument.

Following Sraffa, let us suppose that an extremely simple society is producing just enough wheat and iron to maintain itself. If 400 quarters of wheat (hereafter 400 W) were produced using 280 W and 12 tons of iron (12I) and 1/2 of the annual labour supply (1/2 L) as inputs, while 20I were produced using 120 W, 8I and 1/2 L, then the methods of production and productive consumption can be tabulated as follows:

$$1/2\mathrm{L} + 280\mathrm{W} + 12\mathrm{I} \rightarrow 400\mathrm{W}$$
$$1/2\mathrm{L} + 120\mathrm{W} + 8\mathrm{I} \rightarrow 20\mathrm{I}$$

In order for the process to be repeated the wheat industry must exchange 120 W for 12I. This restores the original distribution of products and enables the process to be repeated.

A three-product model takes us from barter to triangular trade: an *n*-product model to more complex forms of exchange and distribution.

The general formulation of the concept 'production and productive consumption' implicit in Sraffa's analysis is:

$$\text{labour} + \text{things} \rightarrow \text{things}.$$

In other words, the methods of production and productive consumption describe the process of the production of things by means of things and labour. The production by commodities by means of commodities is one historically specific form of these general relations. The emergence of things and labour as commodities presupposes private property and the emergence of a class of proletarians (Marx 1867). This is only one of many social forms that things and labour can take. In tribal economies, for example, things and labour assume the social form of gifts. The social precondition for this to arise is a relatively egalitarian distribution of land between clans. Social data of this kind mean that the principles governing the exchange and distribution of products will vary greatly from economy to economy. In a 'pure' tribal economy, for example, profit maximization is not the central organizing principle of economic life and wages, prices and profits are not be found (Polanyi 1944).

A corollary of this general formulation of production is that 'consumption and consumptive production' can be described as follows:

$$\text{things} + \text{people} \rightarrow \text{people}.$$

In other words 'consumption and consumptive production' describes the methods of production of people by means of people and things.

Neither Marx nor Sraffa analysed these relations which under capitalism would be called the 'household economy' or 'kinship'. However, anthropologists who have studied third world tribal and peasant societies have tended to focus almost exclusively on these relations, a fact, I would suggest, which tells something about the relative importance of production and consumption in capitalist and tribal/peasant societies respectively.

Some indication of what is involved in this concept of consumption can be gleaned by elaborating its meaning in the context of Sraffa's 'extremely simple economy'. Suppose that the iron and wheat were produced by two different households, each household consisting of a father (M), a mother (F), a boy (m) and girl (f). Reproduction of the households, and hence of labour, requires that the children set up new households and produce their own children. Incestuous relations aside, it is clear that the households must exchange children in a way that is analogous to the exchange of wheat for iron discussed above. This can be seen from the following formulation of the relations of consumption and consumptive production for this two-household economy:

$$\begin{aligned} M_1 + F_2 &\rightarrow m_1 + f_1 \\ M_2 + F_1 &\rightarrow m_2 + f_2 \end{aligned}$$

where the subscripts represent the respective households. This particular example is an example of what anthropologists call 'cross cousin' marriage or 'sister exchange'. By tracing the relationships out it will be seen that a man marries his mother's brother's daughter who is also his father's sister's daughter. Take m_1 for example. His father is M_1, his father's sister is F_1, and the latter's daughter is f_2 with whom he will set up a household in the next generation. Tracing the relationships through m_1's mother (F_2) it is obvious that f_2 is also his mother's brother's daughter. Relations of this kind are very important in clan-based societies where a number of households, usually related either matrilineally or patrilineally, occupy a common piece of territory and forbid marriage within the households that make up the clan. In our own society, where the clan has no operational significance, and where marriage is a matter of personal choice rather than a formal arrangement between groups, the political and economic significance of kinship and marriage is relatively unimportant (Gregory 1982).

Every economic analysis of a particular socio-economic form such as 'profits', 'prices' or 'wages' involves, either implicitly or explicitly, a general conceptual framework within which the analysis is carried out. The general model implicit in Quesnay's analysis of 18th-century French agriculture has been elaborated and developed to provide an extremely useful framework not only for the development of a 20th-century theory of the value and distribution of commodities but also for the

analysis of comparative economic systems. By focusing on the circular process of production and reproduction, consumption becomes a dynamic process rather than the dead end of a one way avenue.

See Also

► Economic Anthropology

Bibliography

Gregory, C.A. 1982. *Gifts and commodities*. London: Academic.
Marx, K. 1859. *A contribution to the critique of political economy*. London: Lawrence & Wishart. 1971.
Marx, K. 1867. *Capital*, vol. I. Moscow: Progress.
Polanyi, K. 1944. *The great transformation*. New York: Rinehart.
Quesnay, F. 1759. The 'Tableau Economique'. In *The economics of physiocracy*, ed. R. L. Meek. London: George Allen, 1962.
Sraffa, P. 1960. *Production of commodities by means of commodities*. Cambridge: Cambridge University Press.

Consumption Externalities

Robert H. Frank

Abstract

Consumption externalities occur when consumption by some creates costs or benefits for others. According to Duesenberry's 'relative income hypothesis', spending is influenced by the individual's own standard of living in the recent past and the living standards of others in the present. This hypothesis tracks observed behaviour more closely than Friedman's 'permanent income hypothesis', which assumes that context has no influence on spending. When context is more important for some goods (positional goods) than for others (non-positional goods), positional goods crowd out non-positional goods, causing welfare losses like those that occur when bombs crowd out consumption in military arms races.

Keywords

Bequest motive; Consumption externalities; Friedman, M; Hirsch, F; Marx, K; Permanent income hypothesis; Positional goods; Relative income hypothesis; Revealed preference; Savings; Smith, A; Veblen, T

JEL Classifications

D11

Consumption externalities occur when consumption by some creates external costs or benefits for others. Their recognition by economists dates at least as far back as Adam Smith's discussion of how local consumption standards influence the goods that people consider essential (or 'necessaries', as Smith called them). In the following passage, for example, he described the factors that influence the amount someone must spend on clothing in order to be able appear in public 'without shame':

> By necessaries I understand, not only the commodities which are indispensably necessary for the support of life, but whatever the custom of the country renders it indecent for creditable people, even of the lowest order, to be without. A linen shirt, for example, is, strictly speaking, not a necessary of life. The Greeks and Romans lived, I suppose, very comfortably, though they had no linen. But in the present times, through the greater part of Europe, a creditable day-labourer would be ashamed to appear in publick without a linen shirt, the want of which would be supposed to denote that disgraceful degree of poverty which, it is presumed, no body can well fall into without extreme bad conduct. (Smith 1776, pp. 869–70)

Consumption externalities received only limited attention in Smith's *Wealth of Nations* and only occasional mention by economists during the century that followed its publication. Karl Marx (1847), for example, noted that 'A house may be large or small; as long as the neighboring houses are likewise small, it satisfies all social requirement for a residence. But let there arise next to the little house a palace, and the little house shrinks to a hut.'

It was not until Thorstein Veblen's *The Theory of the Leisure Class* appeared in 1899 that

consumption externalities received their first serious, book-length treatment in economics. Veblen's thesis was that much of consumption is undertaken to signal social position. But although his book is still widely read and cited by scholars in numerous disciplines, its general theme was largely ignored by economists during the 50 years following its publication.

Duesenberry's Relative Income Hypothesis

Interest in this theme was rekindled with the publication of James Duesenberry's *Income, Saving, and the Theory of Consumer Behavior* in 1949. In this volume, Duesenberry offered his 'relative income hypothesis', in which he argued that an individual's spending behaviour is influenced by two important frames of reference – the individual's own standard of living in the recent past and the living standards of others in the present. Thus, in Duesenberry's account, people are subject to both intrapersonal and interpersonal consumption externalities.

His theory attempted to explain three important empirical regularities: (*a*) long- run aggregate savings rates remain roughly constant over time, even in the face of substantial income growth; (*b*) aggregate consumption is much more stable than aggregate income in the short run; and (*c*) individual savings rates rise substantially with income in cross-section data. When Duesenberry's book was first published, individual consumption was generally modelled by economists as a linear function of income with a positive intercept term. This model could accommodate rising savings rates in cross-section data and the stability of consumption over the business cycle, but not the long-run stability of aggregate savings rates.

Duesenberry's hypothesis was hailed as an advance because of its ability to track all three stylized fact patterns. The poor save at lower rates, he argued, because they are more likely to encounter others with desirable goods that are difficult to afford. Moreover, since this will be true no matter how much national income grows, unfavorable comparisons will always occur more frequently for the poor – and hence the absence of any tendency for savings rates to rise with income in the long run.

To explain why consumption is more stable than income in the short run, Duesenberry argued that families compare their living standards not only to those of others around them but also to their own standards from the past. The high consumption level once enjoyed by a formerly prosperous family thus constitutes a frame of reference that makes cutbacks difficult when income falls.

Despite Duesenberry's success in tracking the data, many economists felt uncomfortable with his relative income hypothesis, which to them seemed more like sociology or psychology than economics. The profession was therefore immediately receptive to alternative theories that purported to explain the data without reference to softer disciplines. The most important among these theories was Milton Friedman's permanent income hypothesis, variants of which still dominate today's research on spending.

In hindsight, however, there remain grounds for scepticism about whether Friedman's theory was a real step forward. For example, its fundamental premise – that savings rates are independent of permanent income – has been refuted by numerous careful studies (see, for example, Carroll 1998). Some modern consumption theorists have responded by positing a bequest motive for rich consumers, a move that begs the question of why leaving bequests should entail greater satisfaction for the rich than for the poor.

Another problem is that, contrary to Friedman's assertion that the marginal propensity to consume out of windfall income should be nearly zero, people actually consume such income at almost the same rate as permanent income (Bodkin 1959). To this observation, Friedman (1963) himself responded that consumers appear to have unexpectedly short planning horizons. But if so, then consumption does not really depend primarily on permanent income.

Abundant evidence suggests that context influences evaluations of living standards (see, for example, Veenhoven 1993; Easterlin 1995; Luttmer 2005). In the light of this evidence, it seems fair to say that Duesenberry's hypothesis

not only has been more successful than Friedman's in tracking how people actually spend but also rests on a more realistic model of human nature. And yet the relative income hypothesis is no longer even mentioned in most leading economics textbooks. Its absence appears to signal the profession's continuing reluctance to acknowledge concerns about relative consumption.

Welfare Implications

In traditional economic models, individual utility depends only on absolute consumption. These models lie at the heart of claims that pursuit of individual selfinterest promotes aggregate welfare. In contrast, models that include concerns about relative consumption identify a fundamental conflict between individual and social welfare. This conflict stems from the fact that concerns about relative consumption are stronger in some domains than in others. The disparity gives rise to expenditure arms races focused on 'positional goods' – those for which relative position matters most. The result is to divert resources from 'non-positional goods', causing welfare losses. (The late Fred Hirsch 1976, coined these terms.)

The nature of the misallocation can be made clear with the help of two simple thought experiments. In each, you must choose between two worlds that are identical in every respect except one. The first choice is between world A, in which you will live in a 4,000-square-foot house and others will live in 6,000-square-foot houses; and world B, in which you will live in a 3,000-square-foot house, others in 2,000-square-foot houses. Once you choose, your position on the local housing scale will persist.

If only absolute consumption mattered, A would be clearly better. Yet most people say they would pick B, where their absolute house size is smaller but their relative house size is larger. Even those who say they would pick A seem to recognize why someone might be more satisfied with a 3,000-square-foot house in B than with a substantially larger house in A.

In the second thought experiment, your choice is between world C, in which you would have four weeks a year of vacation time and others would have six weeks; and world D, in which you would have two weeks of vacation, others one week. This time most people pick C, choosing greater absolute vacation time at the expense of lower relative vacation time.

The modal responses in these two thought experiments suggest that housing is a positional good and vacation time a non-positional good. The point is not that absolute house size and relative vacation time are of no concern. Rather, it is that positional concerns weigh more heavily in the first domain than in the second.

When the strength of positional concerns differs across domains, the resulting conflict between individual and social welfare is structurally identical to the one inherent in a military arms race. When deciding how to apportion available resources between domestic consumption and military armaments, each country's valuations are typically more context-dependent in the armaments domain than in the domain of domestic consumption. After all, being less well armed than a rival nation could spell the end of political independence. The familiar result is a mutual escalation of expenditure on armaments that does not enhance security for either nation. Because the extra spending comes at the expense of domestic consumption, its overall effect is to reduce welfare. Note, however, that if each country's valuations were equally context-sensitive in the two domains, there would be no arms race, for in that case the attraction of having more arms than one's rival would be exactly offset by the penalties of having lower relative consumption.

For parallel reasons, the modal responses to the two thought experiments suggest an equilibrium in which people consume too much housing and too little leisure (for a formal demonstration of this result, see Frank 1985a). In contrast, conventional welfare theorems, which assume that individual valuations depend only on absolute consumption, imply optimal allocations of housing and leisure.

In addition to leisure, goods that have been classified as non-positional by various authors include workplace safety, workplace democracy, savings and insurance. And since public goods

are, by definition, available in equal quantities to all consumers, they, too, are inherently non-positional. The general claim is that unregulated market exchange will tend to emphasize the production of positional goods at the expense of these and other non-positional goods (Frank 1985b). Among the policies suggested as remedies for this imbalance have been income and consumption taxes, overtime laws, hours laws for commercial establishments, legal holidays, workplace safety and health regulation, non-waivable workers' rights, and tax-financed savings accounts.

Consumption externalities also have implications for the theory of revealed preference, which says that, if a well-informed individual chooses a risky job that pays $600 a week rather than a safer one that pays only $500, he reveals that the safety increment is worth less than $100 to him. If safety is a non-positional good, however, this inference does not follow, for it ignores the fact that, if all workers exchange safety for increased income, the anticipated increase in relative consumption does not occur. The value that workers assign to safety may thus be revealed as much in the patterns of safety regulation they favour as in the nature of the jobs they choose.

See Also

- ▶ Leisure
- ▶ Time use
- ▶ Veblen, Thorstein Bunde (1857–1929)

Bibliography

Bodkin, R. 1959. Windfall income and consumption. *American Economic Review* 49: 602–614.

Carroll, C. 1998. Why do the rich save so much? In *Does Atlas Shrug: The economic consequences of taxing the rich*, ed. J. Slemrod. New York: Oxford University Press.

Duesenberry, J. 1949. *Income, saving, and the theory of consumer behavior*. Cambridge, MA: Harvard University Press.

Easterlin, R. 1995. Will raising the incomes of all increase the happiness of all? *Journal of Economic Behavior and Organization* 27: 35–47.

Frank, R. 1985a. The demand for unobservable and other nonpositional goods. *American Economic Review* 75: 101–116.

Frank, R. 1985b. *Choosing the right pond*. New York: Oxford University Press.

Friedman, M. 1957. *A theory of the consumption function*. Princeton, NJ: Princeton University Press.

Friedman, M. 1963. Windfalls, the horizon, and related concepts in the permanent income hypothesis. In *Measurement in economics*, ed. C. Christ. Stanford: Stanford University Press.

Hemenway, D., and S. Solnick. 1998. Is more always better? *Journal of Economic Behavior and Organization* 37: 373–383.

Hirsch, F. 1976. *Social limits to growth*. Cambridge, MA: Harvard University Press.

Luttmer, E. 2005. Neighbors as negatives: Relative earnings and well-being. *Quarterly Journal of Economics* 120: 936–1002.

Marx, K. 1847. Relation of wage-labour to capital. In *Wage Labour and Capital.* Marx/Engels Internet Archive, 1999. Online. Available at http://www.marxists.org/archive/marx/works/1847/wage-labour/ch06.htm. Accessed 16 August 2005.

Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations,* ed. R. Campbell and A. Skinner. Oxford: Oxford University Press, 1976.

Veblen, T. 1899. *The theory of the leisure class*. New York: Modern Library.

Veenhoven, R. 1993. *Happiness in nations: Subjective appreciation of life in 56 nations*. Rotterdam: Erasmus University.

Consumption Function

Michael R. Darby

Keynes (1936) introduced the consumption function as the relationship between consumption and income. Although Keynes (pp. 95–6) believed this relationship 'a fairly stable function', substantial shifts in the function were soon observed by empirical workers. Much work in the post-World War II era achieved functional forms by the 1970s which admirers and critics alike could agree were relatively shiftless. Most recent work has considered not functional form but whether or not observed changes in consumption are consistent with models of efficient markets.

The Keynesian Conception

Keynes conceived of the consumption function as relating consumption to disposable income as these are now conventionally measured in the national income accounts. These concepts were basic to the model of *The General Theory* and Keynes was doubtless pedagogically correct to posit a simple relationship which could be refined by future research.

The need for refinement became apparent shortly. In longer time series, consumption seemed to vary around a constant fraction of disposable income. In contrast, consumption functions fitted to depression-era or cross-section data seemed to indicate that this ratio (which Keynes called the average propensity to consume or APC) declined as disposable income rose. In other words, these studies estimated that the derivative of consumption with respect to disposable income (the marginal propensity to consume or MPC) was less than the APC.

Alvin Hansen (1939) among others predicted that a *secular stagnation* would result unless government spending filled this growing gap between output and consumption. When the gap failed to appear, the time was ripe for more sophisticated theories of the relationship between consumption and income. These theories were the earliest and perhaps still most successful resorts to microeconomic foundations for macroeconomics.

Permanent Life-Cycles

In the early 1950s our two dominant models of consumption developed: the permanent-income and life-cycle hypotheses. While these models were once viewed as competing, they can now be seen as complementary with differences in emphasis which serve to illuminate different significant problems. Both models emphasized the distinction between consumer expenditures measured by the national income accounts and pure consumption which was to be explained by optimal allocation of present and future resources over time. The permanent income hypothesis (PIH) stressed stochastic variations in income (and consumption) over time and viewed saving in terms of a bequest motive. The life-cycle hypothesis (LCH) stressed predictable variations in income (and consumption) over the life cycle and viewed saving as resulting from the greater wealth and numbers of younger savers in comparison to older dissavers.

The original published references are to Friedman (1957) for the PIH and Modigliani and Brumberg (1954) for the LCH. Given the delays in NBER publication of Friedman's work which was widely circulated in manuscript form, the two hypotheses are generally regarded as distinct, contemporaneous responses to the described conflict between earlier studies and Simon Kuznets' data on the national income accounts for the 20th century. From the perspective of the monumental careers of the two principal proponents, priority does not seem an issue that need be resolved here.

The PIH relates (pure) consumption to the perpetuity stream that could be consumed forever. The agent is typically regarded as an infinitely lived individual. This represents the underlying notion of a family whose generations are linked by operative transfers from parent to child or vice-versa. Saving arises to equate the ratio of marginal utility of present and future consumption to the marginal rate of transformation implicit in market (real) interest rates. In this way the PIH is said to emphasize the bequest motive for saving.

In contrast, the strict LCH had individuals consuming their entire endowments over their lifetime. Saving was supposed to arise because young workers were more numerous and wealthy (due to technological progress) than the older generation who were dissaving to finance retirement. This provides an avenue by which faster growth can increase saving. Alternatively, as discussed below, factors such as social security which change the extent of mismatch between lifetime consumption and income patterns are predicted to have profound effects on aggregate saving.

These approaches – and their synthesis with inter-generationally linked utility functions – have led to a rich literature quite apart from the consumption function, but those developments are beyond the scope of this essay.

From the point of view of the consumption function per se, the PIH and LCH imply that pure consumption is a fraction (variable in principle but rarely in practice) of wealth or permanent income. Here wealth is inclusive of human as well as non-human capital and permanent income is a (conventionally constant) long-term *ex ante* real interest rate times this wealth. (Note that, contrary to Sargent (1978) and others this wealth is not the discounted present value of expected future income to the extent, as in the PIH, that future income is expected to rise through planned saving.) The empirical estimation of wealth or permanent income became a central issue in the specification of the consumption function.

Friedman proposed a computationally simple estimator of permanent income as a geometrically weighted average of past income. Since on this scheme, permanent income changes – besides normal growth – by a fraction, say β, of the difference between current income and permanent income, Friedman related this scheme to the adaptive-expectations approach recently introduced by his student Phillip Cagan (1956).

Modigliani and his associates proxied normal labour income by current income and the product of this variable and the unemployment rate and attempted to measure non-human wealth by collecting estimates of the national balance sheet at market values. In principle, this method seemed more clearly related to the underlying framework than Friedman's permanent-income proxy, but in practice it suffered several comparative disadvantages: (1) major components of non-human wealth had no market valuation; (2) the wealth estimates were not part of the national income accounts and competing variants were available with substantial delay and at irregular intervals; (3) for forecasting purposes, substantial additional equations were required to forecast (often poorly) future movements in wealth.

Darby (1974) reconciled these empirical measures of wealth by demonstrating that under the PIH, Friedman's geometrically weighted measure could be derived as the constant real interest rate β times a (backward-looking) perpetual inventory of wealth. This β value was estimated as about 0.10 per annum in contrast to higher values such as Friedman's 0.35 per annum. These higher values were explained by biases that arise as data deviate from pure consumption toward expenditures by consumers.

Empirical work on consumption functions has frequently floundered on the use of theories of pure consumption to explain data which are in whole or part consumer expenditures. Both the PIH and LCH were theories of pure consumption. Modigliani and Ando provided one link to consumer expenditures in their MPC model by modelling household investment in durable goods analogously to firm's investment behaviour. Operating in the PIH tradition, Darby (1972, 1974) argued that aggregate transitory income represented a change in wealth, part of which change would be invested in consumers' durable goods. (Darby (1972) in particular argued that because transitory income is received in non-human form, a disproportionate effect on durable-goods purchases may arise during the adjustment process, a result which explains the results of Hayashi 1982.) Darby (1975, 1977–8) later combined pure consumption and durable investment equations to obtain a unified consumer expenditure function which avoided some of the inherent difficulties in dividing consumer expenditures into durable and nondurable portions.

The PIH and LCH thus evolved to explain aggregate consumer expenditures by wealth as a determinant of pure consumption and by *changes* in wealth and other variables which determine household investment in durable goods. The correlation of the determinants of this household investment with short-run (transitory) fluctuations in income explain a MPC which is substantial in magnitude even though substantially below the APC.

This brief development has omitted discussion of alternative views of the consumption function. Perhaps the most notable of these is the view that the substantial value of the MPC reflects liquidity constraints which prevent a substantial share of consumers (measured by wealth and consumption) from following their optimal intertemporal consumption plan. The author of this essay regards these alternative views as providing qualification of the dominant wealthbased view.

Efficient-Market Approaches

Hall (1978) proposed to sidestep Friedman's backward-looking measure of wealth as well as the substantial empirical problems involved in measuring the market value of wealth. Instead, he posed the question of whether or not changes in consumption can be modelled empirically as determined by 'news'. Specifically, the assertion is that if wealth estimates and hence consumption are based on rational expectations, no past information including past changes in consumption or income should affect current changes in consumption.

Hall (1978) answered his question affirmatively, Flavin (1981) dissented, but Hayashi (1982) showed that excess sensitivity of spending to changes in wealth appeared to be confined to consumers' durable goods purchases. Taken as a whole, these studies seem to confirm the basic Friedman–Modigliani conceptions that aggregate consumption as determined by wealth but that it is important to distinguish between consumer expenditures and consumption.

Bequest Versus Life-Cycle Saving

Saying that consumers optimally allocate wealth leaves several important questions unanswered: Do consumers have operative linkages in utility functions across generations? Are consumers able to see through the veil of government to the ultimate production possibilities faced by society? If the first of these questions is answered affirmatively, transfer programmes such as social security which change the life-cycle pattern of income receipts will not affect aggregate consumption and saving. (The representative infinitely lived individual does not care whether he or she pays social security taxes which are refunded as equal benefits. Intergenerational transfers can be adjusted so that this representation is acceptable where utility functions are linked across generations.) If the second question is also answered affirmatively, then Ricardian equivalence holds (it is irrelevant whether government taxes or borrows) and the relevant income concept for the aggregate consumption function is net national product less government spending for goods and services.

Feldstein (1974) claimed that aggregate saving had been significantly reduced by the US social security programme. As pointed out by Barro (1978), this effect would not arise with intergenerationally linked utility functions. Using different methodology, White (1978) and Darby (1979) concluded that life-cycle motives accounted for an at most small fraction of aggregate saving and wealth. Kotlikoff and Summers (1981) relaxed Darby's assumptions on smooth growth of population and labour income without substantially changing the estimates on the range of assets attributable to life-cycle motives. These estimates seem to suggest that intergenerational linkages are indeed very important, as assumed by the PIH. The life-cycle effects highlighted in the LCH would appear more important for analysing cross-sectional data than as determinants of aggregate consumption.

The Ricardian equivalence idea was urged by Barro (1974) and Kochin (1974). It requires a certain suspension of disbelief to assume that bonds and taxes have equivalent effects on consumer behaviour, but the data are not very inconsistent with that notion. Indeed recent studies by Seater (1982) and Kormendi (1983) provide some evidence that Ricardian equivalence is a better working hypothesis than its denial.

Conclusions

The consumption function suggested by Keynes provided a useful challenge to theoretical and empirical economists. The relationship between changes in consumer expenditures and current income has been explained generally in a way which is consistent with microeconomic foundations and which is adequate in a multitude of specifications for most forecasting purposes. (Technical differences among empirical specifications are as large in number as they are uninteresting to the nonspecialist.) For policy analytic purposes, two key questions are outstanding: are life-cycle effects significant in the aggregate,

and do individuals effectively see through government? This author's reading of the evidence suggests answers of no and maybe, but it is hard to put much certainty in any answer unless one starts with dogmatic priors.

The consumption function has faded as a topic of intense research largely because of the success of previous work in achieving a workable consensus. The unsettled issues, however, have crucial policy implications and there is much value yet to be added.

See Also

- ▶ Consumer Expenditure
- ▶ Keynes's General Theory
- ▶ Life Cycle Hypothesis
- ▶ Real Balances
- ▶ Relative Income Hypothesis
- ▶ Wealth Effect

Bibliography

Barro, R.J. 1974. Are government bonds net wealth? *Journal of Political Economy* 82(6): 1095–1117.

Barro, R.J. 1978. *The impact of social security on private saving: Evidence from the U.S. time series*. Washington, DC: American Enterprise Institute.

Cagan, P. 1956. The monetary dynamics of hyperinflation. In *Studies in the quantity theory of money*, ed. M. Friedman. Chicago: University of Chicago Press.

Darby, M.R. 1972. The allocation of transitory income among consumers' assets. *American Economic Review* 62(5): 928–941.

Darby, M.R. 1974. The permanent income theory of consumption: A restatement. *Quarterly Journal of Economics* 88(2): 228–250.

Darby, M.R. 1975. Postwar U.S. consumption, consumer expenditures, and saving. *American Economic Review* 65(2): 217–222.

Darby, M.R. 1977–8. The consumer expenditure function. *Explorations in Economic Research* 4(5): 645–674.

Darby, M.R. 1979. *The effects of social security on income and the capital stock*. Washington, DC: American Enterprise Institute.

Feldstein, M. 1974. Social security, induced retirement, and aggregate capital accumulation. *Journal of Political Economy* 82(5): 905–926.

Flavin, M.A. 1981. The adjustment of consumption to changing expectations about future income. *Journal of Political Economy* 89(5): 974–1009.

Friedman, M. 1957. *A theory of the consumption function*, NBER general series. Vol. 63. Princeton: Princeton University Press.

Hall, R.E. 1978. Stochastic implications of the life cycle – permanent income hypotheses: Theory and evidence. *Journal of Political Economy* 86(6): 971–987.

Hansen, A.H. 1939. Economic progress and declining population growth. *American Economic Review* 29: 1–15.

Hayashi, F. 1982. The permanent income hypothesis: estimation and testing by instrumental variables. *Journal of Political Economy* 90(5): 895–916.

Keynes, J.M. 1936. *The general theory of employment, interest, and money*. New York: Harcourt, Brace, and Co.

Kochin, L.A. 1974. Are future taxes anticipated by consumers? *Journal of Money, Credit, and Banking* 6(3): 385–394.

Kormendi, R.C. 1983. Government debt, government spending, and private sector behavior. *American Economic Review* 73(5): 994–1010.

Kotlikoff, L.J., and L.H. Summers. 1981. The role of intergenerational transfers in aggregate capital accumulation. *Journal of Political Economy* 89(4): 706–732.

Modigliani, F., and R. Brumberg. 1954. Utility analysis and the consumption function: An interpretation of cross-section data. In *Post Keynesian economics*, ed. K.E. Kurihara. New Brunswick: Rutgers University Press.

Sargent, T.J. 1978. Rational expectations, econometric exogeneity, and consumption. *Journal of Political Economy* 86(4): 673–670.

Seater, J.J. 1982. Are future taxes discounted? *Journal of Money, Credit, and Banking* 14(3): 376–389.

White, B.B. 1978. Empirical tests of the life cycle hypothesis. *American Economic Review* 68(4): 647–660.

Consumption Sets

Peter Newman

JEL Classifications
E2

The idea of consumption sets was introduced into general equilibrium theory in July 1954 in Arrow and Debreu (1954, pp. 268–9) and Debreu (1954, p. 588), the name itself appearing only in the latter paper. Later expositions were given by Debreu (1959) and Arrow and Hahn (1971) and a more

general discussion by Koopmans (1957, Essay 1). Although there have been several articles concerned with nonconvex consumption sets (e.g. Yamazaki 1978), in more recent years their role in general equilibrium theory has been muted, especially in approaches that use global analysis (see for example, Mas-Colell 1985, p. 69). Such sets play no role in partial equilibrium theories of consumer's demand, even in such modern treatments as Deaton and Muellbauer (1980). Since general equilibrium theory prides itself on precision and rigour (e.g. Debreu 1959, p. x), it is odd that on close examination the meaning of consumption sets becomes unclear. Indeed, three quite different meanings can be distinguished within the various definitions presented in the literature. These are given below (in each case the containing set is the commodity space, usually R^n): M1 The consumption set C1 is that subset on which the individual's preferences are defined. M2 The consumption set C2 is that subset delimited by a natural bound on the individual's supply of labour services, i.e. 24 hours a day. M3 The consumption set C3 is the subset of all those bundles, the consumption of any one of which would permit the individual to survive. Each definition in the literature can (but here will not) be classified according to which of these meanings it includes. In probably the best known of them (Debreu 1959, ch. 4), the consumption set appears to be the intersection of all three subsets C1–C3. M1 is plain. After all, preferences have to be defined on *some* proper subset of the commodity space, since the whole space includes bundles with some inadmissibly negative coordinates. M2 is also reasonable, although a full treatment of heterogeneous labour services does raise problems for what is meant by an Arrow–Debreu 'commodity' (see for example, that of Arrow–Hahn 1971, pp. 75–6). It is M3 that gives real difficulty, both in itself and in relation to the others.

First, there is little reason to expect either C1 or C3 to be a subset of the other, and so still less to expect M1 and M3 to define the same set. No individual would have any problem in preferring one bundle, the consumption of which would ensure her survival, to a second bundle, the consumption of which would result in her death by starvation. However, she might well prefer the second bundle to a third, whose consumption would cause her to die from thirst (the representation of such preferences by a real-valued utility function might pose problems, but that is another matter). On the other hand, the same individual might not be able to rank in order of preference two bundles each of which contains exotic food and drink, even though fully assured that the consumption of either bundle would allow her to survive.

More importantly, M3 implicitly introduces *consumption* activities, the actual eating and drinking and sheltering that are essential to survival. Such activities constitute what are sometimes called, by analogy with production, the consumption technology. Some partial equilibrium models, such as 'the new home economics' and the theory of characteristics, have treated aspects of such technologies but so far general equilibrium theory has not. In particular, Arrow–Debreu theory has not done so. As a consequence (and unlike some forms of the classical 'corn model') it does not give a coherent account of the birth and death of individual persons, any more than it does of the birth and death of individual firms (*see* general equilibrium). Hence the third meaning M3, which in effect presumes that the model contains such an account when it does not, is hard to interpret. One major difficulty of interpretation arises with the Slater-like condition that each individual's endowment of goods and services, valued at the competitive prices p^*, should be strictly greater than inf $\{(x, p^*): x \in C\}$, where $< .,. >$ denotes inner product and C is 'the' consumption set (see cost minimization and utility maximization). This condition is important in proofs of existence of competitive equilibrium, to ensure for example that the budget correspondence is continuous, or that a compensated equilibrium is a competitive equilibrium. It is itself guaranteed by assumptions (discussed by McKenzie 1981, pp. 821–5) on the relations between 'individual' consumption sets and the aggregate production set.

If C is taken to contain C3 then the assumptions just referred to imply that every consumer

survives in every competitive equilibrium, not merely for one period but over the whole (finite) Arrow–Debreu span. This is a breathtaking assertion of fact which recalls irresistibly Hicks's wry observation: 'Pure economics has a remarkable way of producing rabbits out of a hat – apparently *a priori* propositions which apparently refer to reality. It is fascinating to try to discover how the rabbits got in' (1939, p. 23).

On the other hand if C is taken to be C1, then the assumptions take on a purely technical (and so less objectionable) aspect, whose role is essentially to ensure that the system stays within the (relative) interior of the sets concerned and so displays appropriate continuity. But then there is no presumption that individual agents survive in a competitive equilibrium, even for one period (cf. Robinson 1962, p. 3). The multi-period versions of the Arrow–Debreu model are then at risk, since individuals disappear and take their labour service endowments with them. This should not come as a surprise – the problems of time in economics are really too complicated to be overcome simply by adding more dimensions to the one-period model.

Some models that include C3 in C attempt to justify Slater-like conditions directly, on the grounds that ' Not many economies in the present day are so extremely laissez faire as to permit people to starve' (Gale and Mas-Colell 1975, p. 12). This justification clearly fails as long as the behaviour of the public agency whose actions allegedly prevent such starvation is not modelled *explicitly*, like that of the private agents.

It is usually assumed that consumption sets are bounded below, closed and convex. The first two assumptions are innocuous but the third poses issues of a conceptual kind, which spring from difficulties in interpreting the idea of a convex combination $x^t = tx^1 + (1 - t)x^2$ of two bundles x^1 and x^2, where $t\varepsilon[0, 1]$. Consider the example, sometimes used, in which x^1 is a house in London and x^2 a house in Paris. We cannot take seriously the claim that x^t is a house in the Channel, so t cannot refer to distance. An alternative claim that t refers to the proportion of the period that is spent in London could arise from many different finite partitions of the time interval, not all of which need to be ranked equally by the individual. In effect, convexity of the consumption set comes down to the divisibility of consumer goods, an assumption which in the past has proved not such a bad approximation if one is interested mainly in general equilibrium aspects of market demand, and representative rather than actual consumers. Indivisibilities of producer goods are of course much more serious.

See Also

- ▶ Arrow–Debreu Model of General Equilibrium
- ▶ Cost Minimization and Utility Maximization
- ▶ General Equilibrium
- ▶ Indivisibilities

Bibliography

Arrow, K.J., and G. Debreu. 1954. Existence of an equilibrium for a competitive economy. *Econometrica* 22: 265–290.

Arrow, K.J., and F.H. Hahn. 1971. *General competitive analysis*. San Francisco: Holden-Day.

Deaton, A., and J. Muellbauer. 1980. *Economics and consumer behaviour*. Cambridge: Cambridge University Press.

Debreu, G. 1954. Valuation equilibrium and Pareto optimum. *Proceedings of the National Academy of Sciences* 40 (7): 588–592.

Debreu, G. 1959. *Theory of value*, Cowles commission monograph no.17. New York: Wiley.

Gale, D., and A. Mas-Colell. 1975. An equilibrium existence theorem for a general model without ordered preferences. *Journal of Mathematical Economics* 2: 9–15.

Hicks, J.R. 1939. *Value and capital*. Oxford: Clarendon Press.

Koopmans, T.C. 1957. *Three essays on the state of economic science*. New York: McGraw-Hill.

McKenzie, L.W. 1981. The classical theorem on existence of competitive equilibrium. *Econometrica* 49: 819–841.

Mas-Colell, A. 1985. *The theory of general economic equilibrium. A differentiable approach*. Cambridge: Cambridge University Press.

Robinson, J.V. 1962. The basic theory of normal prices. *Quarterly Journal of Economics* 76 (1): 1–20.

Yamazaki, A. 1978. An equilibrium existence theorem without convexity assumptions. *Econometrica* 46: 541–555.

Consumption Taxation

James M. Poterba

Abstract

Whether to tax households based on their income or on their consumption is one of the central and long-standing questions of tax design. Most developed nations rely on a combination of income and consumption taxes to raise revenue. The debate over alternative tax bases involves both philosophical arguments about what constitutes a fair measure of ability to pay and economic arguments about the relative efficiency of different tax bases. Consumption taxes can be implemented in a variety of ways, including value added taxes, retail sales taxes, and savings-exempt income taxes.

Keywords

Capital gains taxation; Consumption taxation; Distortionary taxation; Flat rate tax; Individual Retirement Accounts (USA); Progressive and regressive taxation; Redistribution; Retail sales tax; Savings-Exempt income tax; Tax compliance; Taxation of income; Value added tax

JEL Classifications

H2

Whether household income or household consumption constitutes a better measure of a household's ability to pay taxes, and whether there are substantial efficiency gains to choosing one tax base rather than the other, are two of the central questions of public finance. The debate between advocates of income taxes and advocates of consumption taxes has spanned several centuries. While income has often been viewed as the basis for taxation, and Adam Smith discusses taxation relative to household incomes, Thomas Hobbes, John Stuart Mill and Irving Fisher were all strong proponents of taxing consumption. Consumption tax supporters argue that the amount that an individual draws from the economy's resource pool should determine his or her tax burden. They also point out that an income tax levies a 'double tax' on saving, since saved income is taxed both when it is earned and when the savings yield a return to capital. Kaldor (1955) offers a broad review of the case for consumption taxation. Two notable reports in the late 1970s, one by the Meade Commission (Meade 1978) in the United Kingdom and the other by the staff of the US Treasury Department (1977), outlined the modern cases for consumption taxation and developed specific proposals.

Proponents of income taxation argue that the change in an individual's command over resources between one period and the next is an appropriate measure of 'ability to pay', even if those resources are not immediately consumed. This is the measure of taxable capacity suggested by Robert Murray Haig and Henry Simons: 'Haig–Simons' income. Moreover, they argue that changes in resources should be taxed regardless of whether they arise from labour income or from the returns to past saving.

Income taxes and consumption taxes exhibit different time profiles over the course of a lifetime. When individuals experience a period of retirement before they die, the time profile of tax payments under a consumption tax will fall later in the lifetime than the corresponding payments under an income tax. This is because individuals continue to consume after they stop earning labour income. Retirees under an income tax pay tax only on their capital income, while retirees under a consumption tax pay tax on their total outlays, which are likely to exceed their capital income.

The debate between proponents of consumption taxation and proponents of income taxation concerns whether or not capital income should be taxed. The foregoing philosophical issues notwithstanding, the efficiency cost of taxing capital income has been an active subject of economic research. Chamley (1986) and Judd (1985) argue that the effective distortions from capital taxes cumulate over time as the difference between discounting the future at before-tax and after-

tax interest rates increases with the compounding horizon. They claim that the optimal steady-state capital income tax rate should be zero. However, they also point out that a one-time capital levy is an efficient device for raising revenue. A number of recent studies, described in Auerbach (2006), have examined the robustness of the theoretical claim that the optimal capital tax rate is zero.

Consumption tax proponents, such as Bradford (1980), claim not only that taxing consumption rather than income avoids intertemporal distortions, but also that it solves many of the most difficult measurement and accounting problems associated with income taxation. Under a consumption tax, for example, there would be no distinction between the tax burden on investment projects financed with debt and those financed with equity, or between realized and unrealized capital gains. There would be no need to measure the rate at which long-lived physical assets depreciate, as one must do under an income tax. Income tax proponents respond that some components of consumption may be difficult to measure, and that it is more difficult to tailor consumption taxes than income taxes to achieve redistributive goals.

Formalizing Consumption Taxation Vs. Income Taxation

The essential difference between a consumption tax and an income tax can be illustrated by comparing the lifetime budget constraints that consumers would face under each tax system. An income tax is levied on both labour and capital income. When a household has assets of A_{t-1} at the beginning of period t, these assets earn a pre-tax return r and the household earns labour income of wL where w equals the real wage and L denotes labour supply, the income tax base is $wL + rA_{t-1}$. The income tax not only reduces the after-tax real wage but also lowers the after-tax return to saving. In a life-cycle model in which a household lives for T periods and in which there is no inflation, the life-cycle budget constraint with an income tax is

$$\sum_{t=1}^{T} C_t/(1+r(1-\tau))^t = \sum_{t=1}^{T} (1-\tau)w_t L_t/(1+r(1-\tau))^t + A_0 \quad (1)$$

In this expression, C denotes real consumption spending, and A_0 is the household's initial wealth endowment.

In contrast, the life-cycle budget constraint with a consumption tax levied at rate θ is

$$\sum_{t=1}^{T} (1+\theta)C_t/(1+r)^t = \sum_{t=1}^{T} w_t L_t/(1+r)^t + A_0 \quad (2)$$

The discount rate in this case is the pre-tax return. The consumption tax levied on outlays in each period is equivalent to a tax on labour income *and* the household's initial endowment. If $(1 - v) = 1/(1 + \theta)$, then Eq. 2 can be rewritten as.

$$\sum_{t=1}^{T} C_t/(1+r)^t = \sum_{t=1}^{T} (1-v)w_t L_t/(1+r)^t + (1-v)A_0 \quad (3)$$

The timing of tax payments under the 'wage-and-endowment tax' in (3) is different from that under the consumption outlays tax in (2), but the present value of taxes and the effects on economic incentives are the same under the two systems. The tax on initial endowment is an essential component of this equivalence: a wage tax alone is *not* equivalent to a consumption tax because initial assets escape taxation when only wages are taxed.

The current tax system in most developed nations is a hybrid structure, reflecting some elements of income taxation but also embodying components of a consumption tax. This is most apparent in nations that rely on both an income tax and a consumption tax, such as a value added tax, for a substantial share of government revenue. Even within many income tax systems, however, there are provisions that move toward an income

tax-consumption tax hybrid. In the United States, for example, capital income that accrues in employer-provided pension plans and in a variety of taxpayer-directed retirement saving accounts, such as Individual Retirement Accounts (IRAs), is excluded from income taxation. Some types of capital income are taxed at rates below the top statutory tax rates on wage income. Realized capital gains have often been taxed at preferential rates, and in some cases dividend income to households is also subject to reduced rates of tax. There is substantial variation in tax structures across nations, but the principle of allowing some tax reduction on capital income is widespread. This makes it difficult to assess where any particular nation's tax system falls on the spectrum between an income tax and a consumption tax.

Types of Consumption Taxes

In practice, there are many ways to implement a consumption tax. Two, the retail sales tax and the value added tax, are widely used in practice. Both are examples of indirect consumption taxes, because they are levied without any reference to the consumer's identity. Direct consumption taxes, in contrast, are levied on households by computing their total consumption. In contrast to indirect consumption taxes, direct consumption taxes can be levied at progressive rates. While direct consumption taxes have never been used as the primary revenue source in any nation, they have been actively debated in the policy reform literature. Tax structures that closely resemble direct consumption taxes have been adopted as components of existing tax systems. The two most widely discussed direct consumption tax options are the savings-exempt income tax and the 'X-tax,' a combination of a cash-flow tax on business income and a household wage tax.

A *retail sales tax* (RST) is the simplest consumption tax. It is collected by retailers at the point of final sale, and it corresponds directly to the tax on consumption spending described in Eq. 2 above. In 2006, 44 of the 50 US states levied some form of sales tax, with rates typically between four and seven per cent. There is little experience with RSTs above ten per cent. One unresolved question with regard to proposals that call for significantly higher RSTs is whether the difficulty of monitoring all points of purchase would lead to substantial problems of tax evasion.

A *value added tax* (VAT) is a very common form of consumption tax. Virtually all developed nations with the exception of the United States levy some form of VAT, with rates ranging up to 25 per cent in Denmark, Norway and Sweden. The VAT is collected from businesses on the difference between the gross value of their sales and the cost of any inputs that they purchase from other entities that have already paid VAT.

To illustrate the operation of VAT, consider a bakery that produces and sells bread for $100. The baker's input costs are $30 for flour and $65 for an employee. The bakery earns a $5 profit. If flour is purchased from another firm that has already paid VAT, then the bakery's VAT liability equals $70 times the VAT tax rate, since its value added equals its sales of $100 minus input purchases that have already paid VAT, or $30. Wages are *not* deducted from sales when computing value added. Although the VAT is collected in stages from all firms in a production chain, it is equivalent to an RST at the same rate. One attractive feature of the VAT is that downstream firms, such as the baker in this example, help ensure VAT compliance by upstream firms that supply intermediate goods. In this example if the flour seller cannot provide documentation for its VAT payment, the baker will face tax on value added of $100. Thus the baker has an incentive, all else equal, to purchase inputs from suppliers who pay VAT.

Ebrill et al. (2001) offer a comprehensive discussion of VAT implementation issues and summarize experience with the VAT in both developed and developing nations. The VAT accounts for a substantial share of revenue in most industrialized nations. The treatment of international transactions has proven a source of difficulty in some nations, since exporting firms are typically granted a rebate for their VAT payments. Some tax evasion schemes involve exporting goods to qualify for the rebate and re-importing the same goods without paying VAT on the import. The

taxation of financial services also proves challenging under the VAT.

A *savings-exempt income tax* (SEIT) is a consumption tax that is built on an income tax model. For those who are familiar with an income tax system, it provides a way of shifting to a consumption tax without drastic administrative changes in the tax system. The Nunn-Domenici 'USA Tax', introduced in the US Senate in the mid-1990s and analysed in Ginsburg (1995), was a based on this type of consumption tax.

Under the SEIT, the tax base is income less saving. To prevent taxpayers from simply claiming high levels of saving and thereby avoiding tax liability, saving must be documented in the form of a contribution to a 'qualified account'. Income earned on assets held in the qualified account is not taxed, but withdrawals from the qualified account are included in the tax base. Thus a taxpayer who earns $50,000 and contributes $5,000 would be taxed on $45,000 in the contribution period. If, some years later, when earnings equal $25,000, the taxpayer withdraws $10,000 from the qualified account, she would be taxed on $35,000.

Even though the SEIT taxes the earnings that have accrued on the contributions to the qualified account when the funds are withdrawn from this account, the return on capital is untaxed in this setting. Taxing accumulated capital income when the proceeds are withdrawn is *not* equivalent to taxing capital income as it accrues: this is the reason Individual Retirement Accounts, 401(*k*) plans and other tax-deferred saving programmes provide an incentive for personal saving. When capital income is taxed as it accrues, the value of earning one dollar, paying tax on it at rate τ, and then investing it for T periods at a pretax rate of return r but with an accrual tax rate τ, is $(1-\tau)(1+(1-\tau)r)^T$. In contrast, if the initial earnings are excluded from taxation, there is no taxation of accruing capital income, and withdrawals are taxed at 100τ per cent, then the value after T periods is $(1-\tau)(1+r)^T$. The qualified account approach eliminates the tax burden on the 'inside build up' of capital assets.

One of the key challenges in implementing a SEIT is avoiding the wholesale reallocation of existing wealth into 'qualified accounts' at the time the SEIT is adopted. Such transfers could sharply reduce tax collections, but, since they involve previously accumulated assets, they would not translate into marginal incentives for new saving. If it were possible to inventory the assets of each taxpayer when the SEIT was implemented, this would make it possible to design regulations to limit the transfer problem. Absent such information on previously accumulated wealth, however, transfers of pre-existing wealth into qualified accounts are likely to prove a difficult implementation issue for the savings-exempt income tax.

An *X-tax* combines a cash flow tax on businesses, much like a VAT with a deduction for wages, with a household-level tax on wage income. The X-tax and its relatives are descended from proposals in the US Treasury Department's (1977) report on fundamental tax reform. Bradford (1986) discusses several plans of this type, and one widely discussed variant was developed by Hall and Rabushka (1995). The X-tax has greater flexibility than a VAT for achieving distributional goals, since the household level tax can include progressive rates or transfers to low-earning households. This illustrates the distributional flexibility of direct rather than indirect consumption taxes. If the household tax is a flat rate tax on wages at the same rate as the corporate cash flow tax, then the X-tax is equivalent to a VAT or an RST. When the rates are different, then the X-tax becomes a combination of a VAT and an additional tax or subsidy on labour income. The cash flow nature of the business tax eliminates the need to measure depreciation, since firms can claim an immediate deduction – expensing – for purchases of capital goods.

In practice, neither the RST nor the VAT is implemented strictly along the principles described above. Proposals for both the SEIT and the X-tax also include additional features that often introduce efficiency costs that would not arise in 'textbook' versions of these taxes. The RST, for example, typically exempts some goods and services. Expenditures on food, medical care and clothing are often excluded from the tax base, thereby achieving a more progressive distribution of tax burdens while creating

distortions between various classes of consumption goods. The VAT is often implemented at different rates on different goods, with exemptions for some goods, creating the same distortionary effects. Because both the savings-exempt income tax and the X-tax require households to file tax returns, they are prone to modification to allow deductions for some expenditure categories, such as mortgage interest or health insurance premiums. While neither of these consumption tax plans has been tried in practice, they probably would be influenced by the same political pressures that have generated a wide array of tax expenditures in the current income tax code.

Efficiency Gains from Replacing an Income Tax with a Consumption Tax

Income taxes create two distortions: one between the before-tax and the after-tax real product wage, which distorts the labour–leisure margin, and one between the beforetax and the after-tax real rate of return to saving. The latter distorts the lifetime allocation of consumption relative to the pattern that would be chosen if the return to delaying consumption equalled the economy's pre-tax marginal product of capital. Shifting from an income tax to a consumption tax eliminates the second distortion. The key analytical issue in evaluating the welfare consequences of replacing an income tax with a consumption tax is therefore measuring the efficiency costs associated with the taxation of saving and investment. This efficiency cost depends on the underlying structure of consumer preferences. The interest elasticity of saving is often invoked as a summary measure of the key preference parameters. When changes in after-tax returns induce only modest changes in household saving, the efficiency gain from switching from an income tax to a consumption tax will be smaller than when the interest elasticity of saving is large.

Auerbach and Kotlikoff (1987) use a dynamic general equilibrium model, including a realistic treatment of household life-cycle income and consumption streams, to evaluate the efficiency gains from replacing an income tax with a consumption tax. Their results suggest that for a given revenue requirement, the steady-state capital stock is larger with a consumption tax than with an income tax. This translates into higher steady-state per capita utility under the consumption tax than the income tax.

The steady-state comparison is not the only consideration when evaluating two alternative tax systems, however. It is possible to design tax reforms that raise steady-state welfare but cause welfare losses in the transition from an initial equilibrium to the new steady state. The trade-off between short-run and long-run policy effects depends on the policymaker's discount rate and in calibrated general equilibrium models it is possible to compute the present discounted value of the gains and losses to the cohorts alive at different dates.

Transition from One Tax Regime to Another

Focusing on the present value of welfare gains and losses draws attention to the transitional rules that govern the switch from one tax system, say an income tax, to another, such as a consumption tax. These transition rules can determine whether a policy reform represents a net gain or a net loss relative to continuation of the initial income tax regime. Altig et al. (2001) illustrate this important point using a more elaborate version of the model developed in Auerbach and Kotlikoff (1987). They find that if the tax basis of existing assets is extinguished when the income tax is replaced by a consumption tax, so that depreciation allowances are no longer claimed after the reform, and if investors who accumulated savings under the income tax regime do not receive any relief from the consumption tax burden they will face when they draw down their assets, then the efficiency gains from adopting a consumption tax may be as large as five per cent of national income.

'Grandfathering' existing assets sharply reduces these efficiency gains, because it reduces the base of the consumption tax and requires

higher tax rates to satisfy a given revenue constraint. This results in greater distortions on the labour–leisure margin. Designing transition relief that participants in the political process will view as fair, without forgoing most of the efficiency gains from a stark consumption tax transition, is likely to be one of the greatest challenges in any consumption-oriented tax reform.

See Also

- ▶ Tax Expenditures
- ▶ Taxation of Income
- ▶ Value-Added Tax

Bibliography

Altig, D., A.J. Auerbach, L.J. Kotlikoff, K.A. Smetters, and J. Walliser. 2001. Simulating fundamental tax reform in the United States. *American Economic Review* 91: 574–595.

Auerbach, A.J. 2006. The choice between income and consumption taxes: A primer. Working Paper No. 12307. Cambridge, MA: NBER.

Auerbach, A.J., and L.J. Kotlikoff. 1987. *Dynamic fiscal policy*. Cambridge: Cambridge University Press.

Bradford, D. 1980. The case for a personal consumption tax. In *What should be taxed? Income or expenditure*, ed. J. Pechman. Washington, DC: Brookings Institution.

Bradford, D. 1986. *Untangling the income tax*. Cambridge, MA: Harvard University Press.

Chamley, C. 1986. Optimal taxation of capital income in general equilibrium with infinite lives. *Econometrica* 54: 607–622.

Ebrill, L., M. Keen, J.-P. Bodin, and V. Summers. 2001. *The modern VAT*. Washington, DC: International Monetary Fund.

Ginsburg, M. 1995. Some thoughts on working, saving, and consuming in Nunn-Domenici's tax world. *National Tax Journal* 48: 585–602.

Hall, R., and A. Rabushka. 1995. *The flat tax*. 2nd ed. Stanford: Hoover Institution Press.

Judd, K.L. 1985. Redistributive taxation in a simple perfect foresight model. *Journal of Public Economics* 29: 59–83.

Kaldor, N. 1955. *An expenditure tax*. London: George Allen & Unwin.

Meade, J.E. 1978. *The structure and reform of direct taxation*. London: Allen & Unwin.

US Treasury Department. 1977. *Blueprints for basic tax reform*. Washington, DC: US Treasury Department.

Consumption-Based Asset Pricing Models (Empirical Performance)

Fatih Guvenen and Hanno Lustig

C

Abstract

Asset pricing is a branch of financial economics that is rich in puzzles and anomalies – that is, stylized empirical facts not easily explained by the canonical asset pricing models. These range from the equity premium puzzle and the risk-free rate puzzle to the fact that stock returns are highly predictable. This article discusses different consumption-based asset pricing models that have been developed to resolve these puzzles, and it evaluates their empirical performance.

Keywords

Capital asset pricing model; Consumption-based asset pricing models; Elasticity of intertemporal substitution; Equity premium puzzle; External habit; Habit formation; Heteroskedasticity; Imperfect risk sharing; Incomplete markets; Precautionary savings; Real business cycles; Recursive preferences; Representative agent; Risk aversion; Risk-free rate puzzle; Stochastic discount factor

JEL Classifications

D4; D10; G12

The aim of consumption-based asset pricing models is to explain a number of important and puzzling features of asset returns using standard economic theory. Perhaps the best-known challenge for these models is the *equity premium puzzle.* Let us start from the Euler equations for stock and bond choice, and let us assume that both of these Euler equations hold with equality. If agents have constant relative risk aversion (CRRA) preferences and if returns and consumption growth are jointly log-normal, then the

Sharpe ratio (that is, the equity premium per unit of risk) can be decomposed as:

$$\frac{E(R^e)}{std(R^e)} \approx \alpha \times std(\Delta c) \times corr(\Delta c, R^e), \quad (1)$$

where R^e is the excess return on stocks over bonds, α is the relative risk aversion (RRA) parameter, and Δc denotes log consumption growth. The equity premium is about 6 per cent per year in the US data with a standard deviation of 15 per cent, producing a Sharpe ratio ($E(R^e)/std(R^e)$) of 0.4. Mehra and Prescott (1985) used the construct of a representative agent who consumes the aggregate endowment stream. Constantinides (1982), Rubinstein (1974) and Wilson (1968) derived aggregation results that rely on either complete markets or the absence of idiosyncratic income risk. By appealing to these aggregation results, Mehra and Prescott could substitute *per-capita* consumption growth into (1). This series has a standard deviation of less than 2 per cent in the post-war US data, and a low correlation with stock returns – less than 0.25 by most estimates. Substituting these values into the expression above implies a lower bound for the relative risk aversion coefficient of 80, which is implausibly high judging by its implications for an individual's choices in other settings. In other words, we need extremely high risk aversion to rationalize the observed equity premium, and that is the puzzle. Furthermore, even if one is willing to accept such a high coefficient of risk aversion, this choice creates different puzzles itself – a point first noted by Weil (1989).

To understand Weil's 'risk-free rate puzzle', first note that the Euler equation for the risk-free asset choice can be linearized to obtain:

$$E\left[R^f\right] \approx -\ln \beta + \alpha E(\Delta c) - \frac{\alpha^2}{2} var(\Delta c). \quad (2)$$

Let us assume a positive time discount rate ($\beta < 1$), and an average consumption growth rate of 1.5 per cent per year. Let us also abstract from uncertainty for the moment. Then a risk aversion of 40 would imply an implausibly high interest rate of nearly 60 per cent per year simply because these households are extremely unwilling to substitute consumption over time. As a result, they desire a flat consumption profile and, therefore, would like to transfer resources from the future to today. But since this is not feasible in an endowment economy, the equilibrium risk-free rate needs to be very high to discourage this type of consumption smoothing and make individuals willing to consume their endowment every period.

The last term in (2) captures the precautionary savings motive, which becomes active in the presence of uncertainty. For very high levels of risk aversion, this effect dominates the intertemporal substitution effect, and an increase in the RRA coefficient *reduces* the risk-free rate. Epstein and Zin (1989) developed a class of recursive preferences that disentangles the inverse of the elasticity of intertemporal substitution from the coefficient of risk aversion. As discussed below, these preferences allow one to make progress on the equity premium puzzle without running into the risk-free rate puzzle.

Against the backdrop of Mehra and Prescott's benchmark model, subsequent papers that attempt to resolve these puzzles can be categorized according to whether they modify (i) the preferences, (ii) the endowment process, or (iii) the market and asset structure. We discuss each of these approaches in turn.

The Utility Function

Recursive Preferences

In the case of CRRA utility, the stochastic discount factor (SDF) has the following form: $M_{t,t+1} = \beta(C_{t+1}/C_t)^{-\alpha}$, where C denotes the level of consumption. A drawback of this specification is that it restricts the elasticity of intertemporal substitution (EIS) to be the reciprocal of the RRA parameter when in fact these two parameters capture conceptually distinct aspects of individuals' preferences. Building on work by Kreps and Porteus (1978), Epstein and Zin (1989) and Weil (1989) introduced 'recursive preferences' (also called 'non-expected utility'):

$$U_t = \left[(1-\beta)C_t^{\rho} + \beta E_t\left(U_{t+1}^{1-\alpha}\right)^{\rho/1-\alpha}\right]^{1/\rho}, \quad (3)$$

where α is still the RRA parameter, but now the EIS is captured by a separate parameter: $1/(1-\rho)$. In this case, the SDF is given by:

$$M_{t,t+1} = \left[\beta\left(\frac{C_{t+1}}{C_t}\right)^{\rho-1}\right]^{\gamma}\left(\frac{1}{R_t^M}\right)^{1-\gamma},$$

where $\gamma = \alpha/\rho$, and R_t^M is the total return on the investors' wealth portfolio (including human capital which must be tradable for this representation to be derived; see Epstein and Zin 1989 and Weil 1989). An appealing feature of this SDF is that it combines two components that are each central to separate asset pricing theories: in particular, the SDF is a geometric average of consumption growth and the market return, where the latter is the relevant SDF in the standard capital asset pricing model (CAPM). Moreover, when $\alpha = 0$ (logarithmic risk preferences), then the CAPM emerges as a special case whereas $\alpha = \rho$ reduces it to the standard case of expected utility (see Epstein and Zin 1989; Campbell 2000).

In addition, this preference specification is flexible enough to allow a choice of a coefficient of relative risk aversion that is high enough to match the equity premium without being forced to accept a very low EIS. The low EIS is responsible for the risk-free rate puzzle, as explained above. Bansal and Yaron (2004) exploit this agent's concern for long-run consumption risk by introducing a small predictable component in consumption growth.

Habit Formation and Catching-Up with the Joneses

Another approach, pioneered by Sundaresan (1989), Abel (1990) and Constantinides (1990), starts from the following specification of the investor's preferences over consumption streams C_t:

$$U_t = \frac{(C_t - X_t)^{1-\alpha}}{1-\alpha}$$

where X_t is some function of either (i) the individual's own past consumption or (ii) the past consumption of a reference group, such as an individual's peers, neighbours, or the population as a whole. Abel's specification features the ratio of C_t to X_t instead of the level difference. The first approach allows an individual's marginal utility to depend on her own past consumption history. This is commonly referred to as habit formation, endogenous habit, or internal habit. The second interpretation allows an individual's utility to depend on her status *relative* to her peers, neighbours or the population as a whole. This is referred to as catching-up with the Joneses or as external habit. These preference specifications amplify the effect of consumption growth shocks on the marginal utility growth of investors, in turn generating a high equity premium.

A particularly successful version of the catching-up-with-the-Joneses specification was developed by Campbell and Cochrane (1999) (henceforth CC) who choose the sensitivity of X to consumption growth shocks to match the conditional and unconditional moments of returns. In the baseline CC model, aggregate consumption and dividend growth are i.i.d. over time. Menzly et al. (2004) introduce additional cash flow dynamics to explain the time series and cross-section of stock returns, while Santos and Veronesi (2005) emphasize the importance of labour income share variation to understand time variation in risk premia. Wachter (2002) applies a version of the CC model to the term structure, while Verdelhan (2004) uses the same model to explain the forward premium puzzle.

Looks Like Habit

Several recent papers have proposed models with standard preferences (such as CRRA) but consider economic environments that give rise to SDFs similar to those resulting from external habit preferences (such as the one used in CC). Examples include work by Piazzesi et al. (2007) who introduce housing services consumption into this framework, and by Yogo (2006) who considers durable consumption broadly defined, building on earlier work by Dunn and Singleton (1986) and Eichenbaum and Hansen (1990). Finally, Guvenen (2005) studies a model with limited stock market participation and shows that

while the asset pricing implications of his model are similar to those in CC, the implications for macroeconomic questions (such as policy analysis, and so on) are quite different.

Additional Arguments in the Utility Function

The models discussed so far assume that investors only derive utility from non-durable consumption. In exchange economy models (in which the consumption process is exogenous) this is equivalent to assuming that non-durable consumption enters the utility function in a separable manner. Some recent papers explicitly model the utility flow from housing consumption (in a non-separable manner), and find that such an extension improves the asset pricing performance (see Grossman and Laroque 1990; Piazzesi et al. 2007; Flavin and Yamashita 2002). Similarly, a labour–leisure choice was introduced by Boldrin et al. (2001) and Danthine and Donaldson (2002), in a representative agent framework, and by Uhlig (2006) in an incomplete markets framework. However, these authors find that this extension negatively affects the performance of asset pricing models, because it allows households to smooth their marginal utility by adjusting on the labour–leisure margin. As a result, one needs to introduce additional – typically labour market – frictions to counteract this new smoothing opportunity.

Consumption Dynamics

In consumption-based asset pricing models, it is common to assume that aggregate consumption growth is i.i.d. over time, because the evidence for consumption growth predictability in the data is weak. In the i.i.d. case, the conditional market price of risk, which can be approximated by the conditional standard deviation of the log SDF, $\sigma_t(logM_{t,t+1}) = \alpha \times \sigma_t(\Delta c)$, is constant. Therefore, these models cannot generate any time variation in risk premia on equity or any other asset.

In the context of a standard representative agent model, Kandel and Stambaugh (1990) generate time-variation in risk premia by introducing heteroskedasticity in aggregate consumption growth. Bansal and Yaron (2004) deviate from the i.i.d. assumption by introducing a small predictable component in consumption growth that is statistically hard to detect. This long-run component increases the market price of consumption risk. In addition, they add some time variation in the size of the long-run risk component. Colacito and Croce (2005) show these long-run risk models can reconcile the low volatility of exchange rate changes with the large market price of risk. Finally, Longstaff and Piazzesi (2002) argue that corporate earnings are much more risky than aggregate consumption growth, and that this can account for a large share of the equity premium puzzle.

Production Economy Models

These asset pricing puzzles have also attracted a lot attention from macroeconomists because the same basic framework used in Mehra and Prescott (1985) also forms the backbone of the Kydland and Prescott (1982) model and the subsequent real business cycle literature. Therefore, understanding why individuals dislike risk in financial markets could help shed light on individuals' perceptions of macro risk and consumption fluctuations, which are key issues for macroeconomic policy. However, macroeconomists are also interested in the determination of quantities, such as output, investment and consumption, making the exchange economy framework unsuitable for their purposes. Therefore, macroeconomists replace the exogenous endowment stream with the endogenous equilibrium consumption process generated by a standard neoclassical production economy that faces technology shocks. One of the first findings of this approach, summarized in Rouwenhorst (1995), is that resolving the equity premium puzzle in a production economy is far more challenging than in an exchange economy, because this endogenous consumption process becomes too smooth if one increases risk aversion. As a result, one needs to resort to real frictions such as large adjustment costs in Jermann's (1998) model. Furthermore, and as noted above, allowing for an endogenous labour supply choice, as is common in macroeconomic analysis, gives

consumers another margin to smooth marginal utility and further reduces the equity premium. Boldrin et al. (2001) and Uhlig (2006) have successfully introduced labour market frictions to effectively shut down this channel.

Market and Asset Structure

The aggregation results we appeal to in order to use a representative agent in asset pricing depend on market completeness. A natural question is to ask what happens if some of these markets are shut down.

Incomplete Markets

In an attempt to resolve the equity premium puzzle, *uninsurable* idiosyncratic income risk has been introduced into consumption-based asset pricing models by Aiyagari and Gertler (1991), Telmer (1993), Lucas (1994), Heaton and Lucas (1996), Krusell and Smith (1997) and Marcet and Singleton (1999), among others. Their main results, obtained numerically for a range of parameter values, suggest that the impact of uninsurable labour income risk on the equity premium is small, because agents manage to smooth consumption quite well by trading a risk-free bond. In fact, Levine and Zame (2002) show that under general conditions the equilibrium allocations and prices in incomplete market economies converge to the complete market counterparts as households become more patient, rendering the incompleteness moot.

So when does imperfect risk sharing matter? Mankiw (1986) derives a sufficient condition for imperfect risk sharing to increase the equity risk premium: the cross-sectional variance of consumption growth needs to increase when returns are low (that is, in recessions). Constantinides and Duffie (1996) embed this counter-cyclical cross-sectional variance mechanism in a general equilibrium model. Grossman and Shiller (1982) show that the Mankiw-Constantinides-Duffie (MCD) mechanism breaks down in continuous-time diffusion models, because the cross-sectional variance of consumption growth is deterministic.

Discussion of Other Models

Rietz (1988) was the first to argue that countries like the United States may simply have been very lucky. Hence, the observed history of the US economy may understate the actual probability of economic disasters, such as the Great Depression (at least as perceived by investors). In this case, the volatility of the SDF may be significantly higher than the one estimated from historical time series. As a result, investors will shun stocks and demand a much higher equity premium to hold them. One difficulty with this explanation is that many economic disasters also result in governments reneging on their debt obligations. Barro (2006) extends Rietz's framework by distinguishing between two types of disasters – those that only affect the stock market and those that affect all asset markets – and explores the empirical implications of this mechanism in recent work.

See Also

- ▶ Capital Asset Pricing Model
- ▶ Consumption-Based Asset Pricing Models (Theory)
- ▶ Elasticity of Intertemporal Substitution
- ▶ Incomplete Markets
- ▶ Recursive Preferences

Bibliography

Abel, A.B.. 1990. Asset prices under habit formation and catching up with the Jones. *American Economic Review* 80: 38–42.

Aiyagari, S.R., and M. Gertler. 1991. Asset returns with transaction costs and uninsured individual risk. *Journal of Monetary Economics* 27: 311–331.

Bansal, R., and A. Yaron. 2004. Risks for the long run: A potential resolution of asset prizing puzzles. *Journal of Finance* 59: 1481–1509.

Barro, R. 2006. Rare disasters and asset markets in the twentieth century. *Quarterly Journal. of Economics* 121: 823–866.

Boldrin, M., L. Christiano, and J. Fisher. 2001. Habit persistence, asset returns, and the business cycle. *American Economic Review* 91: 149–166.

Campbell, J.Y. 2000. Asset pricing at the millennium. *Journal of Finance* 55: 1515–1567.

Campbell, J.Y., and J.H. Cochrane. 1999. By force of habit: A consumption-based explanation of aggregate stock market behavior. *Journal of Political Economy* 107: 205–251.

Colacito, R., and M. Croce. 2005. Risks for the long-run and the real exchange rate. Working paper, New York University.

Constantinides, G.M. 1982. Intertemporal asset pricing with heterogeneous consumers and without demand aggregation. *Journal of Business* 55: 253–267.

Constantinides, G.M. 1990. Habit-formation: A resolution of the equity premium puzzle. *Journal of Political Economy* 98: 519–543.

Constantinides, G.M., and D. Duffie. 1996. Asset pricing with heterogeneous consumers. *Journal of Political Economy* 104: 219–240.

Danthine, J.-P., and J.B. Donaldson. 2002. Labour relations and asset returns. *Review of Economic Studies* 69: 41–64.

Dunn, K., and K. Singleton. 1986. Modeling the term structure of interest rates under nonseparable utility and durability of goods. *Journal of Financial Economics* 17: 769–799.

Eichenbaum, M., and L.P. Hansen. 1990. Estimating models with intertemporal substitution using aggregate time series data. *Journal of Business and Economic Statistics* 8: 53–69.

Epstein, L.G., and S. Zin. 1989. Substitution, risk aversion and the temporal behavior of consumption and asset returns: A theoretical framework. *Econometrica* 57: 937–969.

Flavin, M., and T. Yamashita. 2002. Owner-occupied housing and the composition of the house-hold portfolio. *American Economic Review* 79: 345–362.

Grossman, S., and G. Laroque. 1990. Asset pricing and optimal portfolio choice in the presence of illiquid durable consumption goods. *Econometrica* 58: 25–51.

Grossman, S., and R. Shiller. 1982. Consumption correlatedness and risk measurement in economies with non-traded assets and heterogeneous information. *Journal of Financial Economics* 10: 195–210.

Guvenen, F. 2005. A parsimonious macroeconomic model for asset pricing: Habit formation or cross-sectional heterogeneity? Working paper, University of Texas at Austin.

Hansen, L.P., and K. Singleton. 1983. Stochastic consumption, risk aversion, and the temporal behavior of asset returns. *Journal of Political Economy* 91: 249–265.

Heaton, J., and D. Lucas. 1996. Evaluating the effects of incomplete markets on risk sharing and asset pricing. *Journal of Political Economy* 104: 668–712.

Jermann, U. 1998. Asset pricing in production economies. *Journal of Monetary Economics*: 257–275.

Kandel, S., and R.F. Stambaugh. 1990. Expectations and volatility of consumption and asset returns. *Review of Financial Studies* 3: 207–232.

Kreps, D., and E.L. Porteus. 1978. Temporal resolution of uncertainty and dynamic choice theory. *Econometrica* 46: 185–200.

Krusell, P., and J.A. Smith. 1997. Income and wealth heterogeneity, portfolio selection, and equilibrium asset returns. *Macroeconomic Dynamics* 1: 387–422.

Kydland, Finn E., and Edward C. Prescott. 1982. Time to build and aggregate fluctuations. *Econometrica* 50: 1350–1370.

Levine, D., and W. Zame. 2002. Does market incompleteness matter? *Econometrica* 70: 1805–1839.

Longstaff, F., and M. Piazzesi. 2002. Corporate earnings and the equity premium. Working paper, UCLA Anderson School.

Lucas, D. 1994. Asset pricing with unidiversifiable income risk and short sales constraints: Deepening the equity premium puzzle. *Journal of Monetary Economics* 34: 325–341.

Mankiw, G.N. 1986. The equity premium and the concentration of aggregate shocks. *Journal of Financial Economics* 17: 211–219.

Marcet, A., and K. Singleton. 1999. Equilibrium asset prices and savings of heterogeneous agents in the presence of incomplete markets and portfolio constraints. *Macroeconomic Dynamics* 3: 243–277.

Mehra, R., and E. Prescott. 1985. The equity premium: A puzzle. *Journal of Monetary Economics* 15: 145–161.

Menzly, L., T. Santos, and P. Veronesi. 2004. Understanding predictability. *Journal of Political Economy* 112: 1–47.

Piazzesi, M., M. Schneider, and S. Tuzel. 2007. Housing, consumption, and asset pricing. *Journal of Financial Economics* 83: 531–569.

Rietz, T.A. 1988. The equity risk premium: A solution? *Journal of Monetary Economics* 22: 117–131.

Rouwenhorst, G. 1995. Asset pricing implications of equilibrium business cycle models. In *Frontiers of business cycle research*, ed. T.F. Cooley. Princeton: Princeton University Press.

Rubinstein, M. 1974. An aggregation theorem for security markets. *Journal of Financial Economics* 1: 225–244.

Santos, J., and P. Veronesi. 2005. Labor income and predictable stock returns. *Review of Financial Studies* 19: 1–43.

Sundaresan, S. 1989. Intertemporally dependent preferences and the volatility of consumption and wealth. *Review of Financial Studies* 2: 73–89.

Telmer, C. 1993. Asset-pricing puzzles and incomplete markets. *Journal of Finance* 48: 1803–1832.

Uhlig, H. 2006. Macroeconomics and asset prices: Some mutual implications. Working paper, Humboldt University.

Verdelhan, A. 2004. Habit-based explanation of the exchange rate risk premium. Working paper, Boston University.

Wachter, J. 2002. Habit formation and returns on bonds and stocks. Unpublished paper, Stern School of Business, New York University.

Weil, P. 1989. The equity premium puzzle and the risk-free rate puzzle. *Journal of Monetary Economics* 24: 401–424.

Wilson, R. 1968. The theory of syndicates. *Econometrica* 36: 119–132.
Yogo, M. 2006. A consumption-based explanation of the cross-section of expected stock returns. *Journal of Finance* 61: 539–580.

Consumption-Based Asset Pricing Models (Theory)

Fatih Guvenen and Hanno Lustig

Abstract

The essential element in modern asset pricing theory is a positive random variable called 'the stochastic discount factor' (SDF). This object allows one to price any payoff stream. Its existence is implied by the absence of arbitrage opportunities. Consumption-based asset pricing models link the SDF to the marginal utility growth of investors – and in turn to observable economic variables – and in doing so they provide empirical content to asset pricing theory. This article discusses this class of models.

Keywords

Consumption-based asset pricing models; Equity premium puzzle; Euler equations; Sharpe ratio; Stochastic discount factor

JEL Classifications

D4; D10; G12

Consumption-based asset pricing models study the pricing of payoff streams using the covariance of these payoffs with the marginal utility growth of investors.

The central component of a consumption-based asset pricing model is the Euler equation, which imposes restrictions on the covariance between asset returns and the marginal utility growth of investors. An easy and intuitive way to derive this equation is by using a variational argument. Suppose that the optimal consumption path of investor i is given by $\{C_t^i\}_{t=0}^T$ where T is possibly infinite. Suppose further that an asset j is available with a return $R_{t,t+1}^j$ between periods t and $t+1$, and the investor is not facing a binding portfolio constraint with respect to this asset. Then a feasible strategy is to reduce consumption at time t by a small amount ε, invest it in asset j, and consume the proceeds, $C_{t+1}^i + \varepsilon R_{t+1}^j$, in the next period. Assuming a time-separable utility function, with the one-period felicity function denoted by U and a time discount factor of β, this strategy changes the investors' expected lifetime utility by $-U_c(C_t^i, X_t)\varepsilon + E_t\left[\beta U_c\left(C_{i+1}^i \varepsilon R_{t+1}^j\right)\right]$, where E_t is the mathematical conditional expectation operator; X represents the arguments of the utility function other than consumption; and U_c denotes the partial derivative with respect to consumption. The optimality of the original sequence implies that this strategy cannot be profitable for any amount ε and any asset available. Setting this gain to zero and rearranging yields the Euler equation:

$$E_t\left[M_{t,t+1}R_{t,t+1}^j\right] = 1 \qquad \text{where}$$
$$\mathrm{M}_{t,t+1} = \beta\frac{U_c\left(C_{t+1}^i, X_{t+1}\right)}{U_c\left(C_t^i, X_t\right)}. \tag{1}$$

This Euler equation was first derived by Rubinstein (1976) and Lucas (1978) in discrete time, and by Breeden (1979) in continuous time. While this class of models can in principle be used to study a broad variety of assets, this article will focus on stocks and short-term bonds, which have received the greatest attention in the consumption-based asset pricing literature.

In the case of a one-period discount bond with gross return $R_{t,t+1}^f = 1/P_t^f$ – a bond that costs P_t^f dollars today and pays off 1 dollar tomorrow – the Euler equation can be rewritten as

$$P_t^f = E_t\left[M_{t,t+1}\right]. \tag{2}$$

Similarly, when the asset is a stock with ex-dividend price P_t^s and dividend payment D_t, the Euler equation can be rearranged to read $P_t^s = E_t\left[M_{t,t+1}\left(P_{t+1}^s + D_{t+1}\right)\right]$. By forward substitution this equation yields:

$$P_t^s = E_t\left[\sum_{s=1}^{\infty} M_{t,t+s} D_{t+s}\right], \tag{3}$$

which determines the price of a share of equity as the value of all future dividends it entitles discounted by the SDF.

Lucas (1978) and Mehra and Prescott (1985) used a representative-agent endowment economy structure in which the dividend stream, $\{D_t\}_{t=1}^{\infty}$, is exogenously produced by a 'tree'. Furthermore, these dividends are assumed to be perishable ('fruit'), so in equilibrium the price of equity (in the tree) adjusts to the point where the representative agent is *willing* to consume all available dividends: $C_t = D_t$. Substituting this condition into the expression for M in Eq. (1), and then using M in Eqs. (2) and (3) shows that the price of this stock and that of the one-period bond are entirely determined by the stochastic process for D_t together with the functional form for U (we ignore X_t for now).

Hansen and Singleton (1983) tested the representative agent's Euler equation on US consumption data, and found that the model was rejected. In a famous paper, Mehra and Prescott (1985) showed that when one chooses the properties of C_t to match the moments of aggregate consumption in the data ('calibrate the model to data'), the equity premium $E\left(R_{t+1}^s - R_t^f\right)$ generated by the model was about 60 *times* smaller than that observed in the historical US data. This 'equity premium puzzle' has generated enormous interest and led to the development of a wide range of consumption-based asset pricing models in an attempt to resolve it. For further discussion of the empirical performance of these models, see consumption-based asset pricing models (empirical performance).

An alternative way to explain the hurdles these models face is by deriving an empirical lower bound on the volatility of the stochastic discount factor (SDF). Subtracting the Euler equation for bond returns from the one for stock returns yields: $E\left[M_{t,t+1}\left(R_{t+1}^s - R_t^f\right)\right] = 0$. Noting that the left-hand side of this condition can be rewritten as $Cov\left(M_{t,t+1}\left(R_{t+1}^s - R_t^f\right)\right) + E(M_{t,t+1})E(R_{t+1}\ s - R_t^f)$, some simple manipulations yield the following key decomposition:

$$\frac{E\left(R_{t+1}^s - R_t^f\right)}{\sigma\left(R_{t+1}^s - R_t^f\right)} = -\frac{\sigma\left(M_{t,t+1}\right)}{E\left(M_{t,t+1}\right)} corr \left(M_{t,t+1}, R_{t+1}^s - R_t^f\right), \tag{4}$$

where $\sigma(\cdot)$ denotes the standard deviation. Observing that the correlation term is bounded from above in absolute value by 1, we get

$$\frac{E\left(R_{t+1}^s - R_t^f\right)}{\sigma\left(R_{t+1}^s - R_t^f\right)} \le \frac{\sigma\left(M_{t,t+1}\right)}{E\left(M_{t,t+1}\right)}. \tag{5}$$

The left-hand side of this inequality is the 'Sharpe ratio' – the (expected) excess return demanded by investors per unit (standard deviation) of risk they bear – which averages about 0.40 in annual US data. The right-hand side is called the 'market price of risk' or the 'maximum Sharpe ratio'. This inequality bound implies that a consumption-based model must be able to generate an SDF with a coefficient of variation (standard deviation normalized by mean) of at least 40 per cent to be consistent with the Sharpe ratio observed in the data. This observation – developed by Shiller (1982) and further generalized by Hansen and Jagannathan (1991) – provides a 'volatility bounds' test for potential candidate models. As discussed in consumption-based asset pricing models (empirical performance), the majority of plausibly calibrated asset pricing models fail this test.

When the investor faces a binding borrowing constraint, she cannot increase her consumption today by reducing the holdings of asset j. As a result, her marginal utility today will remain higher than the value implied by the equality condition in (1), and the Euler condition for that asset will instead be an inequality: $E_t\left[M_{t,t+1} R_{t,t+1}^j\right] < 1$. This relaxes the lower bound on the volatility of the SDF derived in Eq. (5) (cf. Luttmer 1996).

To develop further implications of consumption-based models it is necessary to

impose additional structure on $M_{t,t+1}$, which requires being more specific about (1) the functional form and the arguments of the utility function; (2) the stochastic properties of variables affecting marginal utility (that is, consumption, leisure, and so on); and (3) the market structure. The latter determines whether an appropriate aggregation theorem holds (which happens for example when markets are complete), in which case C_i can be replaced with aggregate consumption. Therefore, consumption-based models can be broadly categorized based on the assumptions they make along these three dimensions. These different models are discussed in consumption-based asset pricing models (empirical performance).

Another feature of asset markets that has received much attention in the literature concerns the high volatility of stock prices. For example, the standard deviation of the log price/dividend (P/D) ratio of stocks is about 40 per cent per annum in the US data. In a world with a constant SDF (as would be the case with risk-neutral investors), it is impossible to rationalize this high volatility with the relatively low variability of the underlying dividend stream (LeRoy and Porter 1981; Shiller 1981). Let p_t denote the log price, d_t denote the log dividend, and r_t denote the log stock return. Using a first-order approximation, Campbell and Shiller (1988) show that the log P/D ratio can be decomposed as follows:

$$p_t - d_t = constant + E_t \sum_{j=1}^{\infty} \rho^{j-1} \left[\Delta d_{t+j} - r_{t+j} \right]$$

with $\rho = \exp(\overline{pd})/1(1+\exp(\overline{pd}))$ and $\overline{pd}$ denotes the average log P/D ratio. The first term in the square brackets is referred to as the cash flow component, and the second part is referred to as the discount rate component. This decomposition implies that the variance of the log P/D ratio can be stated as:

$$var(p_t - d_t) cov\left(p_t - d_t, \sum_{j=1}^{\infty} \rho^{j-1} \Delta d_{t+j}\right) - cov\left(p_t - d_t, \sum_{j=1}^{\infty} \rho^{j-1} r_{t+j}\right).$$

This expression shows that the P/D ratio moves only because it predicts future returns on stocks or because it predicts future dividend growth. In the data, most of the volatility in P/D ratio is due to news about future expected returns ('discount rates'), not due to future dividend growth ('cash flows') (Campbell 1991; Cochrane 1991). There is a large literature that documents the predictability of stock returns over longer holding periods, starting with work by Campbell and Shiller (1988, 1998), Poterba and Summers (1986) and Fama and French (1988, 1989). Other variables that predict returns include the spread between long and short bonds (Fama and French 1989) and the T-bill rate (Lamont 1998). More recently, more attention has been paid to macroeconomic variables that predict returns, most notably in the work by Lettau and Ludvigson (2001a) who document that the consumption/wealth ratio is a powerful predictor of stock returns.

So, the volatility of P/D ratio implies that excess returns on stocks are highly predictable. In other words, expected excess returns change a lot over time, even per unit of risk. We use the conditional version of the expression in (4) to understand the implications of this finding:

$$\frac{E_t\left(R^s_{t+1} - R^f_t\right)}{\sigma\left(R^s_{t+1} - R^f_t\right)} = \frac{\sigma_t\left(M_{t,t+1}\right)}{E_t\left(M_{t,t+1}\right)} corr_t\left(M_{t,t+1}, R^s_{t+1} - R^f_t\right), \tag{6}$$

where σ_t denotes the conditional standard deviation. Good models need to produce a lot of time variation in the right-hand side of (6) and this happens mostly through variation in the conditional market price of risk (first term). This is an upper bound on the conditional Sharpe ratio. (See also Lettau and Ludvigson 2001b, on how to measure variation in the conditional Sharpe ratio.) Another test of consumptionbased asset pricing models is whether they are able to generate as much predictability as found in the data. Examples of early models that match the variation in the conditional market price of risk include Kandel and Stambaugh (1990), Campbell and

Cochrane (2000) and Barberis et al. (2001). More recent work includes the work by Santos and Veronesi (2005), Menzly et al. (2004), Piazzesi et al. (2007), Guvenen (2005), Lustig and Van Nieuwerburgh (2005, 2006) and Bansal and Yaron (2004). These models are discussed in detail in consumption-based asset pricing models (empirical performance).

See Also

- ▶ Consumption-Based Asset Pricing Models (Empirical Performance)
- ▶ Euler Equations

Bibliography

Bansal, R., and A. Yaron. 2004. Risks for the long-run: A potential resolution of asset pricing puzzles. *Journal of Finance* 59: 1481–1509.

Barberis, N., M. Huang, and T. Santos. 2001. Prospect theory and asset prices. *Quarterly Journal of Economics* 116: 1–53.

Breeden, D.T. 1979. An intertemporal asset pricing model with stochastic consumption and investment opportunities. *Journal of Financial Economics* 7: 265–296.

Campbell, J.Y. 1991. A variance decomposition for stock returns. *Economic Journal* 101: 157–179.

Campbell, J.Y., and J.H. Cochrane. 2000. Explaining the poor performance of consumption-based asset pricing models. *Journal of Finance* 55: 2863–2878.

Campbell, J.Y., and R.J. Shiller. 1988. The dividend-price ratio and expectations of future dividends and discount factors. *Review of Financial Studies* 1: 195–227.

Campbell, J.Y., and R.J. Shiller. 1998. Stock prices, earnings and expected dividends. *Journal of Finance* 43: 661–676.

Cochrane, J.H. 1991. Explaining the variance of price-dividend ratios. *Review of Financial Studies* 5: 243–280.

Fama, E.F., and K.R. French. 1988. Dividend yields and expected stock returns. *Journal of Financial Economics* 22: 3–27.

Fama, E.F., and K.R. French. 1989. Business conditions and expected returns on stocks and bonds. *Journal of Financial Economics* 25: 23–49.

Guvenen, F. 2005. *A parsimonious macroeconomic model for asset pricing: Habit formation or cross-sectional heterogeneity?* Working paper, University of Texas at Austin.

Hansen, L.P., and R. Jagannathan. 1991. Implications of security markets data for models of dynamic economies. *Journal of Political Economy* 99: 252–262.

Hansen, L.P., and K. Singleton. 1983. Stochastic consumption, risk aversion, and the temporal behavior of asset returns. *Journal of Political Economy* 91: 249–265.

Kandel, S., and R.F. Stambaugh. 1990. Expectations and volatility of consumption and asset returns. *Review of Financial Studies* 3: 207–232.

Lamont, O. 1998. Earnings and expected returns. *Journal of Finance* 53: 1563–1587.

LeRoy, S.F., and R.D. Porter. 1981. The present-value relation: Tests based on implied variance bounds. *Econometrica* 49: 555–574.

Lettau, M., and S.C. Ludvigson. 2001a. Consumption, aggregate wealth and expected stock returns. *Journal of Finance* 56: 815–849.

Lettau, M., and S.C. Ludvigson. 2001b. Measuring and modeling variation in the risk-return tradeoff. In *Handbook of Financial Econometrics*, ed. Y. Ait-Sahalia and L.P. Hansen. Amsterdam: North-Holland.

Lucas, R. 1978. Asset prices in an exchange economy. *Econometrica* 46: 1429–1454.

Lustig, H. and S.V. Nieuwerburgh. 2006. *Can housing collateral explain long-run swings in asset returns?* Working Paper NYU Stern and UCLA.

Lustig, H., and S. van Nieuwerburgh. 2005. Housing collateral, consumption insurance and risk premia: An empirical perspective. *Journal of Finance* 60: 1167–1219.

Luttmer, E. 1996. Asset pricing in economies with frictions. *Econometrica* 64: 1439–1467.

Mehra, R., and E. Prescott. 1985. The equity premium: A puzzle. *Journal of Monetary Economics* 15: 145–161.

Menzly, L., T. Santos, and P. Veronesi. 2004. Understanding predictability. *Journal of Political Economy* 112: 1–47.

Piazzesi, M., M. Schneider, and S. Tuzel. 2007. Housing, consumption, and asset pricing. *Journal of Financial Economics* 83: 531–569.

Poterba, J.M., and L.H. Summers. 1986. The persistence of volatility and stock market fluctuations. *American Economic Review* 76: 1142–1151.

Rubinstein, M. 1976. The valuation of uncertain income streams and the pricing of options. *Bell Journal of Economics* 7: 407–425.

Santos, J., and P. Veronesi. 2005. Labor income and predictable stock returns. *Review of Financial Studies* 19: 1–43.

Shiller, R.J. 1981. The use of volatility measures in assessing market efficiency. *Journal of Finance* 36: 291–304.

Shiller, R.J. 1982. Consumption, assest markets and macroeconomic fluctuations. *Carnegie-Rochester Series on Public Policy* 17: 203–238. North-Holland Publishing Company.

Contemporary Capitalism

William Lazonick

Abstract

The key to understanding 'capitalism' as a mode of resource allocation that generates economic growth is the organization and performance of its most innovative business enterprises. The 'Old Economy business model' that made the United States the world's most powerful nation in the post-Second World War decades came under challenge in the 1970s and 1980s, and the ideology of 'maximizing shareholder value' arose to legitimize a redistribution of income from labour interests to financial interests. The 'New Economy business model' emerged in the 1980s and 1990s to drive the innovation process, contributing, however, to unstable and inequitable economic growth.

Keywords

Acquisitions; Business enterprises; Capitalism; Collective capitalism; Competitive advantage; Conglomerate movement; Contemporary capitalism; Corporations; Creative destruction; Defined-benefit pensions; Developmental state; Dividend yield.; Division of labour; Economic development; Entrepreneurship; Financial commitment; Foreign direct investment; Globalization; Great Depression; Hostile takeovers; Information and communications technology; Innovation; Japan, economics in; Junk bonds; Leveraged buyouts; Lifelong employment; Mergers; Milken, M.; NASDAQ; New Deal; New Economy business model; New York Stock Exchange; Old Economy business model; Organization man; Organizational integration; Outsourcing; Patents; Research and development; Schumpeter, J. A.; Separation of ownership and control; Shareholder value; Social inclusion; Stock options; Stock price volatility; Stock repurchases; Strategic control; Technology; Venture capital

JEL Classifications

P1

What Is 'Capitalism'?

At the beginning of the 21st century, 'capitalism' has triumphed as the dominant system for allocating a society's economic resources. The last time in history in which the persistence of capitalism in the world's most advanced economies was seriously called into question was the Great Depression of the 1930s – a decade during which the unemployment rate in the United States remained at 15 per cent or higher, notwithstanding unprecedented state intervention under the New Deal. It took the Second World War to pull the United States and the world economy out of depression, and in the subsequent decades it took substantial and sustained government spending in the rich economies of North America and western Europe to hold unemployment to acceptable levels.

In the post-war era, the Soviet Union's highly planned economy posed as a possible alternative to capitalism. The purported strength of the Soviet challenge, however, turned out to be based at least as much on Cold War ideology emanating from the United States as on the actual productive power of the Soviet Union and its satellites. By the 1990s the Soviet model had virtually vanished, as Russia itself sought to make the transition to a 'market economy', guided, tragically, by a mythical ideology of how capitalism is supposed to operate, imported from the United States.

Over the same period capitalism entrenched itself in East Asia. During the 1970s and 1980s Japan became a rich economy on the basis of a distinctive model of 'collective capitalism', and in the 1980s and 1990s the East Asian 'Tigers' – Hong Kong, Singapore, South Korea

and Taiwan – closed the gap, each with its own variant of the Japanese model. More recently China and India, with one-third of the world's population, have experienced rapid economic growth, driven by what many would call 'capitalist' institutions. Yet, even as firms cross the globe to access Indian software engineers, and vice versa, India remains a nation with one-third of the world's illiterates. Meanwhile the fact that China, the world's second largest economy since the early 1990s, continues to be guided by an avowedly Communist government raises the question of what 'capitalism' really is.

Defining contemporary capitalism is not merely a question of semantics. If, as has been demonstrated since the mid-20th century, 'capitalism' is a powerful engine of economic growth, we want to know how it functions as a mode of resource allocation and the social conditions under which capitalist growth is not only strong but also stable, and equitable. We also want to know how the institutions of contemporary capitalism that generate growth might be transferred to those parts of the world – first and foremost Africa but also eastern Europe and Latin America as well as parts of the Middle East – that have economically been left behind. Given its pervasiveness and dominance, a depiction of the institutions that define contemporary capitalism is tantamount to a description of the economic world in which perhaps one-half of the world's population now lives and to which much of the other half now aspires.

There is no consensus among economists on the definition of contemporary capitalism. The dominant approach to analysing resource allocation and the economic performance of an advanced economy rests on the notion that a capitalist economy is essentially a market economy that allocates resources to their most productive uses. But what at any time and in any place, the student of economic development asks, explains how those most productive uses come to exist? And why in certain times and places? Fundamental to capitalist growth is 'innovation', the process that generates goods and services that, even with factor prices held constant, are of higher quality and lower cost than those previously available (Lazonick 2006c). Can a theory of capitalism as a market economy comprehend the innovation process?

In the early 20th century a young Joseph Schumpeter asked this question. As a Viennese economics student, Schumpeter was versed in the relatively recent, and increasingly influential, Austrian and Walrasian theories of how, through the equilibrating mechanism of the market, the economy could achieve an 'optimal'allocation of resources across productive uses. Schumpeter's insight was to recognize that such a view of the economic world could not explain economic development. In 1911 Schumpeter wrote *The Theory of Economic Development* (first translated into English in 1934) to argue that entrepreneurial activity that results in innovation – what he called the 'Fundamental Phenomenon of Economic Development' – can disrupt the 'Circular Flow of Economic Life as Conditioned by Given Circumstances' to change the ways in which the economy operates and performs. Without such disruption of equilibrium conditions, the economy would not develop. Over the next four decades Schumpeter sought to elaborate a theory of economic development informed by his own, evolving, understanding of the changing reality of the most advanced capitalist economies.

In particular, Schumpeter sought to understand the role of the business enterprise in advanced capitalist development. By the 1940s he had taken definitive leave of his youthful conceptions of the innovative entrepreneur as an individual actor and innovation as simply 'new combinations' of existing resources. Rather, he saw that powerful business organizations both developed and utilized productive resources to create new technologies and access new markets. The creation of new technologies, moreover, destroyed the commercial viability of old technologies. In *Capitalism, Socialism, and Democracy*, first published in 1942, Schumpeter argued that the process of 'creative destruction' had become embodied in established corporations as 'technological "progress" tends, through systematization and rationalization of research and of management, to become more effective and surefooted', being 'the business of teams of trained specialists who turn out what is

required and make it work in predictable ways' (Schumpeter 1950, pp. 118, 132).

This article takes as its point of departure the proposition, suggested by Schumpeter, that the key to understanding 'capitalism' as a mode of resource allocation that generates economic growth is the organization and performance of its most innovative business enterprises. That is not to say that markets and states are unimportant to the operation, and hence definition, of capitalism. Historically, however, well-functioning markets are outcomes of successful capitalist development. For the individual, markets create the possibility of choosing what to consume and for whom to work, including the prospect of working for oneself. But markets cannot explain the development of the new products and processes that drive the growth of the capitalist economy. The innovation process is uncertain, collective and cumulative (see O'sullivan 2000). The uncertain character of innovation means that investments in innovation require *strategic control* over resource allocation by individuals who have intimate knowledge of the technologies, markets and competitors that an innovative strategy must confront. The collective character of innovation means that the implementation of an innovation strategy requires the *organizational integration* of a hierarchical and functional division of labour into a process of organizational learning. The cumulative character of innovation means that the process requires *financial commitment* until it can generate financial returns. Enterprises, not markets, engage in strategic control, organizational integration and financial commitment (Lazonick 2003).

Nor can one explain innovation by appealing to the notion of the developmental state as its driving force, as has often been done for the East Asian economies. Implicit, and at times explicit, in this view is an acceptance of the ideology that the economic development of the United States is an exemplar of the workings of the market economy. Yet from gun manufacture and interchangeable parts in the first half of the 19th century to the computer revolution and Internet in the late 20th century, as well as railroads, aviation and the life sciences in between, the history of US capitalism is replete with examples of the critical role of the developmental state in allocating resources to the processes of knowledge creation that then provided the foundations for US industrial leadership. Yet, as important as the developmental state has been even in a so-called 'market economy' such as the United States, the allocation of resources to knowledge creation would have been wasted, and would probably never have been made, had it not been for the presence and influence of innovative enterprises that have made use of this knowledge to generate higherquality, lower-cost products than had previously been available.

In this article I focus on the changing role of innovative enterprise in determining resource allocation and economic performance in contemporary capitalism. Space constraints dictate that I confine the analysis of contemporary capitalism to the case of the United States, with the caveat that, even in a highly globalized economy in which one might expect convergence to a common business model, there are almost as many distinctive 'varieties of capitalism' in terms of governance, employment and investment institutions, as there are advanced capitalist nations. The US economy is, however, the world's largest and richest economy. It is also the one in which market ideology is most virulent and the actual mode of resource allocation most misunderstood. Section "The Old Economy business model" of this article provides historical background to understanding contemporary US capitalism by describing the key characteristics of the 'Old Economy business model' (OEBM) that made the United States the world's most powerful nation in the decades after the Second World War. Section "Maximizing shareholder value" analyses the challenges that confronted OEBM in the 1970s and 1980s, and how the ideology of 'maximizing shareholder value' arose to legitimize a redistribution of income from labour interests to financial interests. Section "The New Economy business model" shows how the 'New Economy business model' (NEBM) emerged in the 1980s and 1990s to drive the innovation process, but in ways that have contributed to unstable and inequitable economic growth. Section "Stable

and equitable growth?" concludes with some questions about the future of the US model in a global economy in which many distinctive business models still compete.

The Old Economy Business Model

The United States emerged from the Second World War as the undisputed world leader in GDP per capita, a position that it still retains. With western Europe and Japan still in recovery from the war, the United States was at its peak of dominance in the 1950s on the basis of a highly collective model of capitalism embodied in the managerial corporation, and personified in the concept of the 'organization man' (Whyte 1956). The stereotypical 'organization man' was white, Anglo-Saxon and Protestant, obtained a college education, got a well-paying job with an established company early in his career, and then worked his way up and around the corporate hierarchy over decades of employment, with a substantial 'defined benefit' pension, complete with highly subsidized medical coverage, awaiting him on retirement. The employment stability offered by an established corporation was highly valued, while inter-firm labour mobility was shunned.

'Organization men' rose to top executive positions where, as salaried managers rather than owners, they exercised strategic control. This separation of share ownership and managerial control, which continues to characterize the US industrial corporation, resulted from the widespread distribution among shareholders of the corporation's publicly traded stock. In principle, boards of directors representing the interests of shareholders monitor the decisions of these managers. In practice, incumbent top executives choose the outside directors and are themselves members of the board. Shareholders can challenge management through proposals to the annual general meeting, but over the course of the 20th century a body of law evolved that enables management to exclude shareholder proposals that deal with normal business matters (for example, downsizings) as distinct from social issues (for example, sex discrimination).

The separation of ownership from control has worked effectively to generate innovation when the interests of salaried executives who exercise strategic control have been aligned with those of employees who engage in the development and ensure the utilization of the company's productive resources. In the post-Second World War decades the organizational integration of the capabilities of administrative and technical specialists enabled US firms to develop the world's most competitive systems of mass production. These personnel were products of the US system of higher education, which since the early decades of the century had prepared the labour force to enter employment in bureaucratic organizations.

A distinctive feature of the US business model was the organizational segmentation between these salaried managers, in whose training and experience the corporation made substantial investments, and so-called 'hourly' workers. (Nonsalaried employees were classified as 'hourly', or 'non-exempt', workers because of the stipulation of the National Labor Relations Act that emerged from the New Deal era that required employees who were paid an hourly wage receive 150 per cent of that wage if they worked longer than the normal working hours. The overtime work of salaried personnel is exempt from this provision.) The corporation viewed these operatives, who were typically high-school graduates, as interchangeable commodities in whose capabilities the company had no need to invest, notwithstanding the fact that they often spent their entire working lives with one company. At the same time, these industrial corporations needed reliable even if lowskill workers to tend mass production processes. The combination of dominant product-market positions and union power, which advanced the pay and protected the employment of senior workers, enabled the hourly worker to receive good pay and benefits, including a defined-benefit pension that assumed long-term employment with a single company.

The developmental state played an indispensable role in the innovation process by partially funding the system of higher education as well as, in the forms of research labs, subsidies and

contracts, programmes for technology development in sectors such as aerospace, computers and life sciences. The development of the productive potential of these government investments relied in turn on corporate research capabilities. Retained earnings formed the foundation of committed finance for new corporate investments in innovation. When corporations needed additional investment financing, they issued corporate bonds at favourable rates that reflected the established position of the company as well as its conservative debt–equity ratios. Companies used bank loans almost exclusively for working capital, and made only limited use of the stock market as a source of investment funds.

These social conditions enabled US corporations to grow very large in the post-war decades. The 50 largest US industrial corporations by revenues on the Fortune 500 list averaged 87,070 employees in 1957, 117,393 in 1967, and 119,093 in 1977. These figures do not include employment at AT&T, the regulated telephone monopoly, which in 1971 employed 1,015,000 people, of whom 700,000 were union members with good wages, stable employment and excellent benefits. By the late 1960s and early 1970s increasing numbers of blacks were moving into union jobs in the steel, automobile, electrical equipment, consumer durable and telecommunications industries. The growth of established corporations in these industries in the three decades after the Second World War contributed to a more equal distribution of family income in the US economy.

'Maximizing Shareholder Value'

During the 1970s the US model faltered in the face of Japanese competition. Building on innovative capabilities developed for their home markets during the 1950s and 1960s, Japanese companies gained competitive advantage over US companies in industries such as steel, memory chips, machine tools, electrical machinery, consumer electronics and automobiles. US companies had entered the 1970s as world leaders in these industries. Many US observers attributed the rapid increase in Japanese exports to the United States in the 1970s to Japan's lower wages and longer working hours. By the early 1980s, however, with real wages in Japan continuing to rise, it became clear that Japanese advantage was based on the superior organization of their enterprises, and in particular on a more thoroughgoing integration of participants in the functional and hierarchical divisions of labour for the dual purposes of transforming technologies and accessing new markets. Indeed, during the 1980s Japan exported management practices as well as material goods to the West. From the second half of the 1980s, with the yen strengthening and trade surpluses generating political backlash, Japanese companies made a transition to direct investment in the United States and other advanced economies.

A growing financial orientation of US business that had surfaced in the conglomerate movement of the 1960s undermined the abilities and incentives of established US corporations to respond to the Japanese challenge. To some extent the growth of the US industrial corporation in the post-war decades had been based on strategic investments in new product lines and geographic areas that built on the corporation's existing productive capabilities, and yielded economies of scale and scope. The conglomerate movement, however, saw major corporations invest in scores of *unrelated* businesses, often through mergers and acquisitions, based on the prevailing, but erroneous, ideology that a good corporate executive could manage any type of business, and that conglomeration offered the synergies of superior corporate management. The conglomerate movement failed because it segmented top executives, in positions of strategic control, from the rest of the managerial organization that had to develop and utilize productive resources to sustain the firm's competitive advantage (Lazonick 2004).

In the late 1970s and early 1980s the conglomerates unraveled. In the mid-1970s Michael Milken, a Drexel Burnham investment banker, had created the junk bond market by convincing institutional investors, in search of higher yields in an inflationary era, to hold downgraded corporate securities, many of them 'fallen angels' from

Contemporary Capitalism, Table 1 US corporate stock and bond yields, 1960–2005. Average annual per cent change

	1960–9	1970–9	1980–9	1990–9	2000–5
Real stock yield	**6.63**	**−1.66**	**11.67**	**15.01**	**− 1.87**
Price yield	5.80	1.35	12.91	15.54	− 0.76
Dividend yield	3.19	4.08	4.32	2.47	1.58
Change in CPI	2.36	7.09	5.55	3.00	2.67
Real bond yield	**2.65**	**1.14**	**5.79**	**4.72**	**3.60**

Source: Council of Economic Advisers (2006, Tables B-62, B-73, B-95 and B-96)
Notes: Stock yields are for Standard and Poor's composite index of 500 US corporate stocks (424 of which are, as of 28 March 2006, NYSE). Bond yields are for Moody's Aaa-rated US corporate bonds

unsuccessful conglomeration. By the late 1970s, with the junk-bond market well developed, it became possible to issue new junk bonds to finance leveraged buyouts (LBOs) in which the top managers of a conglomerate division turned it into an independent company to recapture strategic control over resource allocation. By the late 1980s, however, the junk bond had become an instrument for the hostile takeover of entire companies, with KKR's 1989 LBO of RJR Nabisco for $24.5 billion marking the height of what became known as 'the deal decade'.

The ideology that justified hostile takeovers was that the corporation should be run to 'maximize shareholder value' (see Lazonick and O'sullivan 2000). Proponents of shareholder value charged that, either because of opportunism or incompetence, many incumbent corporate managers were making poor allocative decisions. By exercising their influence through the market for corporate control, shareholders could force incumbents to alter their allocative decisions, replace them with those who would maximize shareholder value, or distribute cash to shareholders in the forms of dividends and stock repurchases so that shareholders themselves could, so the argument goes, reallocate the economy's resources to their best alternative uses.

While the hostile takeover movement did not directly threaten high-tech companies (in which the most valuable assets could walk out the door), by the end of the 1980s the top executives of virtually all US industrial corporations had embraced the ideology of maximizing shareholder value and made it their own. By the 1980s executive stock option compensation was a well-established practice. Since in the United States option awards did not require that the company's stock price outperform the stock market or even the stock prices of a group of competitors, those who received these awards could only gain from what, from July 1982 to August 2000, turned out to be the longest stock market boom in US history, with the Dow Jones Industrial Average and the S&P500 Index both rising about 1,300 per cent.

As Table 1 shows, stock-price appreciation drove the extraordinary real stock yields that were sustained over the 1980s and 1990s. The relatively low dividend yields in the 1990s did not reflect stinginess on the part of US corporations; the US corporate payout ratio – the amount of dividends as a percentage of after-tax corporate profits (with inventory evaluation and capital consumption adjustments) – averaged 48 per cent in the 1980s and 57 per cent in the 1990s compared with 39 per cent in the 1960s and 41 per cent in the 1970s. It was just that the rate of increase of stock prices outstripped the rate of increase of dividend payments, thus depressing the dividend yield. The form that the stock yield takes is of significance because investors can capture the dividend yield by holding stocks, whereas they can capture the price yield only by buying and selling stocks. Inherent in high-price yields, therefore, is a volatile stock market.

A volatile stock market benefits those who are compensated in stock options on an annual basis, especially when, as is the case in the United States, options vest as quickly as one year from the date of grant and can be exercised for up to ten years. It has been estimated that, largely because of the gains from exercising stock options, on average the ratio of CEO pay of an S&P500

company to that of a production worker was 42 in 1985, 107 in 1990, 525 in 2000, and 411 in 2005. Top executives took a keen interest in their company's stock price, and in the 1980s and 1990s, in the name of 'maximizing shareholder value', they found ways in which they could use their positions of strategic control over corporate resource allocation to influence it. They could cook the corporate books to boost current earnings, a practice that became widespread in these decades and one for which a few executives have been fined or even jailed. The American Competitiveness and Corporate Accountability Act of 2002, better known as Sarbanes–Oxley, has sought to stem this practice. But quite apart from artificially inflating corporate earnings, top corporate executives also found that downsizing the labour force and repurchasing corporate stock helped to boost a company's stock price, even though these resource allocations did not necessarily improve the company's competitive performance.

The era of corporate downsizing took hold in the recession of 1980–2 when hundreds of thousands of stable, well-paid blue-collar jobs were lost that were never subsequently restored (see Lazonick 2004). It would appear that the blacks who had relatively recently moved into these types of jobs were particularly hard hit; last hired, they tended to be the first fired. The subsequent 'boom' years of the mid-1980s witnessed hundreds of plant closings. In the 'white-collar' recession of the early 1990s tens of thousands of professional, administrative and technical employees found that their jobs had been eliminated, although once again it was blue-collar workers who bore the brunt of the downturn. In 1980 manufacturing employment was 22 per cent of the labour force; by 1990 it had fallen to 17 per cent and by 2001 to 14 per cent. While the employment picture generally became much better during the Internet boom of the last half of the 1990s, job cutting remained a way of life for many major US corporations. According to data on layoff announcements by companies in the United States collected by the recruitment firm, Challenger, Gray and Christmas, announced job cuts averaged just under 550,000 per year for the period 1991–4, 450,000 per year in 1995–7, and 656,000 per year during the boom years 1998–2000.

Meanwhile, from the mid-1980s US corporations began to actively support their stock prices through large-scale stock repurchases. Companies included in the S&P500 in March 2006 distributed more cash to shareholders in repurchases than in dividends in 1997 through 2000 and again in 2004, and just slightly less in 2001 through 2003. Since 1978 net equity issues by US non-financial corporations has been positive in only six of 28 years (1980, 1982, 1983, 1991, 1992, 1993); since the early 1980s US industrial corporations have in aggregate been supplying capital to the stock market rather than vice versa. In 2005 the net flow of cash from non-financial corporations to the stock market was a record $366 billion, 1.42 times in real dollars the previous high in 1998 (Lazonick 2006d).

The New Economy Business Model

On 29 December 1995, AT&T announced that, as part of the process of breaking itself up into three separate companies, it would be cutting 40,000 jobs. AT&T was a company that could trace its origins back to the 1870s, had created the world's most advanced telephone system, was the home of the famous Bell Labs that among many other accomplishments invented the transistor in 1947, and, despite having lost its status as a regulated monopoly in 1984, still employed 308,700 people. Now, however, AT&T became emblematic of the failure of US Old Economy corporations to continue to provide employment opportunities. With campaigning for the 1996 presidential election picking up steam, Patrick Buchanan, a right-wing politician, caught the attention of the media by denouncing the highly paid executives of AT&T and other downsizing corporations as 'corporate hit men'. Fuel was added to the fire by the revelation that, in the name of 'creating shareholder value', Al Dunlap, whom the American public came to know as 'Chainsaw Al', had in 20 months as CEO of Scott Paper devastated the 115-year old company while putting an estimated

Contemporary Capitalism, Table 2 Comparing business models in ICT

	Old economy business model (OEBM)	New economy business model (NEBM)
Strategy, product	Firm growth based on multidivisional structure: multi-product firm	New firm entry into specialized ICT markets; accumulate new capabilities by acquiring (other) young technology firms
Strategy, process	Vertical integration of the value chain; in-house standards and proprietary R&D	Vertical specialization of the value chain; industry technology standards; R&D for cross-licensing and alliances; outsourcing routine work to specialist contract manufacturers and/or offshoring routine work to low-wage nations
Finance	Venture finance from savings, family and business associates; NYSE listing, growth finance from retentions, after dividends, and bond issues	Organized venture capital; early IPO on NASDAQ; retentions with zero dividends; use of own stock as a compensation and combination currency; systematic stock repurchases to support stock price
Organization	Secure employment; 'organization man' (career with one company), industrial union; defined-benefit pension, good medical coverage in employment and retirement	Insecure employment; interfirm mobility of labour, broad-based stock options, nonunion, defined contribution pension, employees bear more burden of medical insurance

$100 million in his own pocket. In March 1996, the *New York Times* ran a seven-part series, later released as a paperback, on 'the downsizing of America' (Lazonick 2004).

By the spring of 1996, however, the furor over corporate downsizing had disappeared. In its place, Americans became enthralled by the prosperity promised by what in the second half of the 1990s came to be called the 'New Economy'. In the United States the previous half-century had seen a massive accumulation of information and communications technology (ICT) capabilities. The development of computer chips from the late 1950s had provided the technological foundation for the microcomputer revolution from the late 1970s, which in turn had provided the technological infrastructure for the Internet boom of the second half of the 1990s. The research funding for this accumulation of ICT capabilities had come mainly from the US government and the research laboratories of established Old Economy high-technology corporations. Each wave of technological innovation, however, created opportunities for the emergence of start-up companies that were to become central to the commercialization of the new technologies.

Although by the mid-1980s the Japanese had outcompeted even the leading US semiconductor firms in the memory chip market, US companies such as Intel, Motorola and Texas Instruments continued to dominate the microprocessor and logic chip markets that drove product innovation in the microelectronics industry (Lazonick 2006a). While Silicon Valley was not the only US location for innovation in this industry, the concentration of semiconductor start-ups in the region from the late 1950s resulted in the emergence by the 1980s of a distinctive mode of combining strategy, finance and organization: the 'New Economy business model' (NEBM) (see Table 2). During the 1990s NEBM spread beyond Silicon Valley start-ups and was adopted successfully by leading Old Economy ICT companies such as Hewlett-Packard and IBM. In the Internet boom of the late 1990s elements of NEBM diffused to other ICT companies, including an Old Economy company such as Lucent Technologies, spun off from AT&T in its 1996 trivestiture, which almost destroyed itself in attempting to adopt the business model. In the 2000s NEBM characterizes the most innovative sectors of the US economy (for the case of biotechnology, see Pisano 2006).

The founders of New Economy firms have typically been scientists and engineers who have gained specialized experience in existing firms, although in some cases they have been university faculty members intent on commercializing their academic knowledge. Some of these entrepreneurs have come from existing Old Economy

companies, where it was often difficult for their new ideas to get internal backing. But New Economy companies themselves have become increasingly important as sources of new entrepreneurs who left their current employers to start new firms. Large numbers of high-tech entrepreneurs in the United States have been foreign-born, coming mainly from India and China (Saxenian 2006).

Typically, the founding entrepreneurs of a New Economy start-up seek committed finance from venture capitalists with whom they share not only ownership of the company but also strategic control. In the 2000s Silicon Valley remains by far the leading location in the United States and the world for venture-backed high-tech start-ups. The region acquired this position from the 1960s as a distinctive venture capital industry emerged out of the opportunities for start-ups created by the microelectronics revolution. Besides sitting on the board of directors of the new company, the venture capitalist generally recruits professional managers, who are given company stock along with stock options, to lead the transformation of the firm from a new venture to a going concern. This stock-based compensation gives these managers a powerful financial incentive to develop the innovative capabilities of the company to the point where it can do an initial public offering (IPO) or private sale to a listed company, thus enabling the start-up's privately held shares to be transformed into publicly traded shares. Both before and after making this transition, their tenure with, and value to, the company depends on their managerial capabilities, not their fractional ownership stakes (Lazonick 2006a).

The stock market speculation of the 'dotcom' era made it all too easy to cash out of a start-up, as many high-tech firms that had not engaged in innovation did IPOs or were sold to established companies. When start-ups do innovate, the key to making the transition from new venture to going concern has been the organizational integration of an expanding body of technical and administrative 'talent'. As Silicon Valley developed from the 1960s, this educated and experienced labour had to be induced to trade secure employment with an Old Economy company for insecure employment with a start-up. To attract these highly mobile people and retain their services, Silicon Valley firms increasingly adopted 'broad-based' employee stock option plans that extended this form of compensation to a large proportion, sometimes all, of the firm's non-executive employees rather than just to top executives. In start-ups, stock options usually served as a partial substitute for cash salaries, and the eventual gains from exercising options were viewed as a substitute for a company-funded pension (Lazonick 2006a).

Again, the underlying stock would become valuable if and when the start-up did an IPO or a private sale to a publicly listed company. Shortening the expected period between the launch of a company and an IPO was the practice of most venture-backed high-tech start-ups of going public on NASDAQ, created in 1971 as an electronic exchange for the over-the-counter markets with less stringent listing requirements than the 'Old Economy' New York Stock Exchange (NYSE). The 1978 cut in the capital gains tax rate to 28 per cent, after it had been raised to 49 per cent just two years before, provided further encouragement to entrepreneurs and venture capitalists to found new companies, and for employees of these companies, rewarded with stock options, to provide the skills and efforts needed to transform new ventures into going concerns. In 1979 the clarification of the 'prudent man' rule as applied to the Employee Retirement Income Security Act (ERISA) of 1974 gave asset managers the green light to allocate a portion of their portfolios to riskier stocks and venture capital funds, and resulted in a flood of new money, especially from pension funds, into the venture capital industry. The American Electronics Association and the National Venture Capital Association, with their strongest and deepest roots in Silicon Valley, were the frontline Washington lobbyists for these regulatory changes.

While institutional money provided capital to NEBM, high-tech labour became more mobile from one firm to another than it had been in the Old Economy. Employee stock options induced this mobility, but what made it possible in terms of the knowledge bases that managers and engineers possessed were *industry* standards, as distinct

from *in-house* standards, that emerged in the various sectors of ICT. In the Old Economy in-house standards promoted the growth of large vertically integrated firms on the basis of proprietary technologies, whereas in the New Economy industry standards encouraged new entry. Nevertheless, as demonstrated by the important cases of Intel and Microsoft in the development of the microcomputer industry, those New Economy firms that dominated in the setting of the industry standards could also grow very large (at the end of fiscal 2005 Intel employed 99,900 and Microsoft 61,000). By establishing industry standards, their growth encouraged rather than discouraged start-ups, which in turn depended on the availability of not only venture capital (which came from many sources besides the formal venture capital industry) but also mobile labour whose knowledge and experience could be easily integrated into the start-up's learning processes.

Of critical importance in setting industry standards in microelectronics was IBM's decision in 1980 to enter what became known as the personal computer (PC) industry with Intel supplying the microprocessor and Microsoft the operating system. At the time IBM controlled about 80 per cent of the computer market, had over 341,000 employees, and, with an explicit system of 'lifelong employment', trumpeted the fact that since 1921 it had not terminated an employee involuntarily. Yet between 1990 and 1994 IBM slashed its employment from 374,000 to 220,000. In 1991–3, the company had losses of $16 billion (including more than $8 billion in 1993, at the time the largest annual loss in US corporate history) on total revenues of $192 billion, and encouraged the media to believe that the mass layoffs were necessary to avoid bankruptcy. Yet virtually all of the losses came from 'restructuring' charges, that is, the cost of terminating employees (Lazonick 2006a).

In retrospect, it is clear that these charges were the cost of ridding the company of its 70-year-old system of lifelong employment. The industry standards in ICT, which IBM had played a leading role in establishing, served to reduce the value to the company of older employees with experience accumulated at IBM over the course of their careers and to increase the value of younger employees who may have had experience working for other ICT companies. Explicitly reflecting this change in employment policy, in 1999 IBM announced that it would replace its traditional defined-benefit pension plan, which favoured long-term employees, with a portable 'cash-balances' plan that would be much more attractive to younger employees who did not envisage a lifelong career with IBM. In December 2004, as its employment reached 329,000, IBM announced that new employees would no longer be eligible for the cash-balances pension fund. Instead the company would offer them a defined-contribution pension, with the company matching the employee contribution up to six per cent of his or her compensation.

From the mid-1990s, with the Old Economy commitment to its employees out of the way, IBM adopted all of the elements of NEBM. It shifted out of hardware into services, and outsourced its manufacturing. It became by far the leading patenter in the United States, even as it cut R&D from the ten per cent of sales that prevailed in the 1980s to six per cent of sales since the mid-1990s, this change reflecting an expressed shift to product development and away from basic research. Since the early 1990s IBM has engaged in patenting much less to control proprietary technologies, as had been the case in the past, and much more to gain access through cross-licensing to technologies controlled by other companies and to generate intellectual property revenue ($1.3 billion per year in the 2000s).

As it rid itself of lifelong employment in the early 1990s IBM began to extend stock options, previously reserved for top executives, to a broad base of employees. In 1990 options outstanding were only four per cent of all shares outstanding; in 2005, 15.2 per cent. As for distributions to shareholders, in New Economy fashion, subsequent to its early 1990s restructuring IBM has favoured repurchases over dividends. In 1981–90 IBM's dividends were 48 per cent and repurchases 12 per cent of net income; in 1993–2005 dividends were 15 per cent and repurchases 91 per cent. In an effort to offset dilution of shareholdings as employees exercise

stock options, and more generally to boost its stock price, in 1995–2005 IBM has spent $62.6 billion on stock repurchases. Over the same period the company has spent $56.6 billion on R&D.

As for a New Economy company that, unlike IBM, started out that way, Cisco Systems, which since the late 1990s has controlled about 75 per cent of the Internet router market, is a prime example of the importance, and implications, of broad-based stock options in NEBM compensation. Founded in Silicon Valley in 1984, Cisco grew from about 200 employees at the time of its IPO in 1990 to 40,000 employees during 2000. Throughout its history Cisco has awarded stock options to virtually all of its employees. By the end of fiscal 2000 stock options outstanding accounted for 14 per cent of the company's total stock outstanding; by 2005 that number was 23 per cent. In March 2000, at the peak of the Internet boom, Cisco had the highest market capitalization of any company in the world. Under such conditions its stock options were very lucrative. I have estimated that over the 11 years 1995–2005 (all years for which data are reported refer to fiscal year's end, the last week in July), Cisco employees, totaling about 256,000 employee-years, shared $21.5 billion in gains from exercising stock options, for an average of $84,000 per employee-year. The annual averages per employee ranged from less then $9,000 in 2003 to more than $281,000 in 2000. Of the total amount, the highest paid executives, totaling 57 executive-years, shared $893 million, for an average of $15.7 million per executive-year, with annual averages ranging from $1.3 million in 2003 to $51.3 million in 2000. The annual ratios of average top-executive to average employee gains from exercising stock options ranged from 36:1 in 1997 to 594:1 in 2005 (Lazonick 2006d). Cisco employees have a clear financial interest in the company's stock price, and the company's top executives even more so.

Besides using their own stock as a compensation currency, during the 1990s some New Economy companies grew large by using their stock, instead of cash, to acquire other, smaller and typically younger, New Economy firms in order to gain access to new technologies and markets. Cisco mastered this growth-through-acquisition strategy. From 1993 through 2005 Cisco made 106 acquisitions valued in nominal terms at $46.9 billion, over 80 per cent of which was paid in the company's stock rather than cash. In 1999 and 2000 alone, Cisco did 41 acquisitions at a cost of $26.7 billion with over 99 per cent paid in stock (Carpenter et al. 2003).

At the same time, like many if not most New Economy companies, Cisco conserved cash by paying no dividends. Along with its use of stock as a combination currency, this payout policy enabled Cisco to become a giant company in the 1990s without taking on any long-term debt. Since the bursting of the Internet bubble from mid-2000 through 2005, however, Cisco has spent $27.2 billion repurchasing its own stock to support its sagging stock price. In 2004–5, as it spent $19.3 billion on stock repurchases, Cisco used $8.3 billion in cash – including $6.5 billion of it raised through its first-ever bond issue – to do 24 acquisitions rather than continue to use its stock as an acquisition currency that it would then feel compelled to offset with repurchases. (Cisco's decision to use cash rather than stock for acquisitions was helped by the Financial Accounting Standards Board's 2001 closing of the 'pooling-of-interests' loophole that enabled companies like Cisco that did all-stock acquisitions to record them on their balance sheets at book values, which were generally a small fraction of market values, and thus inflated future earnings. Nevertheless, in 2002 and 2003, with pooling-of-interests accounting outlawed, Cisco still used stock for payment of over 97 per cent of the price of its nine acquisitions.)

The corporate obsession with supporting its stock price through massive stock repurchases has therefore taken hold of companies in the most innovative sectors of the US economy. As further notable examples, for the years 1995–2005 Intel distributed $51.3 billion in repurchases along with $6.0 billion in dividends compared with R&D spending of $38.0 billion, while Microsoft distributed $45.4 billion in repurchases and $38.7 billion in dividends compared with R&D spending of $40.8 billion.

Microsoft's dividends included a one-time payment of $36.1 billion in November 2004.

These companies would argue that R&D spending and stock repurchases are both working toward the same end: to enhance the company's innovative capabilities by, in the case of R&D, generating new knowledge, and in the case of repurchases, competing for high-tech labour capable of transforming that knowledge into innovative products and processes. By boosting stock prices, it is argued, repurchases help to attract, retain and motivate people who choose to work for companies in which they are partially compensated with stock options. In the case of Microsoft the argument has had less weight since July 2003 when the company ended its option programme (although many Microsoft employees still have unexercised options awarded prior to that date). In the 2000s, moreover, the extent and location of the talented labour supplies for which companies like Cisco, IBM, Intel and Microsoft compete have changed dramatically with the rise of India and China (Lazonick 2006b). These dramatically changed labour market conditions for high-tech labour raise serious questions concerning which employees benefit from a company's stock price performance and for how long, and indeed whether established high-tech companies even need to use employee stock options to compete successfully for high-tech labour.

The offshoring to India and China in the 2000s of high value-added jobs of software engineers and computer programmers that it was previously thought could not go abroad represents the latest stage in four decades of the globalization of NEBM. Beginning in the early 1960s Silicon Valley semiconductor companies were among the first to offshore assembly to East Asia, and by the early 1970s virtually every US semiconductor manufacturer had followed suit. When these companies set up plants in places like South Korea, Taiwan, Hong Kong, Singapore and Malaysia, they employed, alongside unskilled and predominantly female assembly labour, indigenous university graduates as managers and engineers. Over time the US companies upgraded their facilities in these locations, and offered more and better employment opportunities for the indigenous well-educated labour force. As a striking example, in 1984 Intel claimed that, of its 8,500 employees outside of the United States (of 26,000 employees worldwide), only 60 were US citizens. This indigenous employment through foreign direct investment encouraged the national governments to increase the level of investment in their already well-developed systems of higher education, thus augmenting the future high-tech labour supply (Lazonick 2006b).

In the 1990s established US ICT companies, led by IBM and Hewlett-Packard, dramatically reduced their employment of production workers by outsourcing manufacturing operations to electronic manufacturing service providers, also known as contract manufacturers (Lazonick 2006c). Indeed, younger companies like Cisco grew rapidly without doing any in-house manufacturing. Initially the contract manufacturers would set up operations or take over existing plants of their customers in the United States. But a key capability of the leading contract manufacturers is to shift production that has become more routine and cost-sensitive to lower wage areas of the world. In the late 1990s and early 2000s the leading contract manufacturers grew at a rapid pace; at the end of 2005 employment at the five largest – Flextronics, Solectron, Sanmina-SCI, Celestica and Jabil Circuit – totalled 260,000. While we do not know the global distribution of this labour force, North America accounts for only an estimated 25 per cent of the sales of these five companies.

Meanwhile, in the 1990s and 2000s hundreds of thousands of foreigners, especially Indians, with college degrees in science and engineering have migrated to the United States for graduate education and work experience (Lazonick 2006b). Many acquired permanent resident (immigrant) status in the United States, as the US government expanded employment-based preferences in the issuance of immigrant visas. For access to US work experience, however, the most important mode of entry for high-tech employees has been on non-immigrant H-1B and L-1 visas. The H-1B programme enables

non-immigrants, the vast majority of whom have at least a bachelor's degree and whose skills are purportedly unavailable in the United States, to work in the United States for up to six years. In the first half of the 2000s about 70 per cent of H-1B visa holders had science or technology degrees, and 40–50 per cent came from India (the next largest national group is from China, at less than ten per cent). The L-1 visa programme permits a company with operations in the United States to transfer foreign employees to the United States to acquire work experience, with no limitation of time. In 2001, there were an estimated 810,000 people on H-1B visas in the United States, and possibly as many on L-1 visas.

Many of these non-immigrant visa holders have continued to work in the United States by obtaining permanent resident status. But most have returned to their native countries with valuable industrial experience that can be used to start new firms and, more typically, to work as technical specialists for indigenous or foreign companies. As a result of both the migration of US companies abroad in search of high-tech labour as well as the migration of foreign high-tech labour to the United States, and then back to their home countries, in the 2000s, to an extent never before imagined, even the best-educated US high-tech employees compete with a truly global labour supply for jobs.

Stable and Equitable Growth?

On 16 March 2005 the Semiconductor Industry Association (SIA) organized a Washington, DC press conference in which it exhorted the US government to step up support for research in the physical sciences, including nanotechnology, to assure the continued technological leadership of the United States. Intel CEO Craig Barrett was there as a SIA spokesperson to warn: 'U.S. leadership in technology is under assault' (*Electronic News* 2005):

> The challenge we face is global in nature and broader in scope than any we have faced in the past. The initial step in responding to this challenge is that America must decide to compete. If we don't compete and win, there will be very serious consequences for our standard of living and national security in the future. . .U.S. leadership in the nanoelectronics era is not guaranteed. It will take a massive, coordinated U.S. research effort involving academia, industry, and state and federal governments to ensure that America continues to be the world leader in information technology.

At the time Barrett was a member of the US National Academy of Sciences Committee on Prospering in the Global Economy in the 21st Century, which delved into deficiencies in the development of science and engineering capabilities in the United States. Notwithstanding his obvious concern about these problems from a public policy perspective, on a radio talk show in February 2006 Barrett (by this time Chairman of Intel) remarked: 'Companies like Intel can do perfectly well in the global marketplace without hiring a single US employee' (wbur.org 2006).

The problem with this statement is not that US workers should have privileged access to jobs at a US-based company like Intel (which still employs half of its almost 100,000 employees in the United States). The problem is that, if a powerful company like Intel is not dependent on US high-tech employees for its future labour force, why should it be concerned about supporting the mass educational infrastructure in the United States needed to develop this future labour force? And what does it mean to say that 'America must decide to compete' if, as I would argue is the case (Lazonick 2006b), the most innovative US corporations have more of an interest in the Malaysian or Indian system of mass education than in the US system?

Since the mid-1970s the US mass education system has been performing poorly in science and mathematics by the standards of both the advanced and many developing economies. Such was much less the case in the three decades or so after the Second World War, when the Old Economy corporation was more dependent upon a labour force that was well-educated at the primary and secondary levels in the United States. This shift in the performance of the mass education

parallels the reversal of post-war progress towards a more equal distribution of income that began about three decades ago. The much less secure employment of most US corporate employees in the shift from OEBM to NEBM would seem to have contributed to this reversal.

Meanwhile in the 2000s the compensation of the CEOs of US corporations has long since passed levels that are at a minimum unseemly and some would say obscene. The 'explosion in CEO pay', which has been discussed in the United States since the mid-1980s, seems to have no limits, especially if, when the corporate stock price falls, it can be once again pumped up or boards of directors can replace the 'lost' stock option income by other forms of remuneration such as salaries, bonuses or restricted stock. The seemingly endless explosions in top-executive pay reflect the obsession of US corporate executives with 'maximizing shareholder value' and, cash flow permitting, disgorging billions upon billions of corporate cash to shareholders in the forms of repurchases and dividends to try to make it happen.

In terms of public policy initiatives, virtually nothing has been done to control top executive pay in the United States. One well-known attempt was misguided. In 1993 President Clinton carried out a campaign promise to control CEO pay by legislating a cap of $1 million on the amount of 'non-performance-related' top executive compensation – salary and bonus – that a corporation could claim as a tax deduction. One perverse result of this law was that companies that were paying CEOs less than $1 million in salary and bonus *raised* these components of CEO pay towards $1 million, which executives now viewed as the government-approved CEO 'minimum wage'. The other perverse result was that companies increased CEO option compensation, for which tax deductions were not in any case being claimed, as an alternative to exceeding the $1 million salary-and-bonus cap.

That having been said, the limits to the gains from stock options, not just for top executives but also for broad bases of the employees of US high-tech corporations, would long ago have been reached if not for the fact that many of these corporations have been in the forefront of innovation. Given the unchallenged sway that the ideology of 'maximizing shareholder value' has over the governance of these corporations, I have no doubt that instability, as reflected in the boom and bust of the stock market in the late 1990s and early 2000s, and inequity, as reflected in the worsening of the distribution of income, will continue to beset the US economy.

Whether US corporations will remain in the forefront of innovation that, by necessity, must underpin long-term economic growth is another matter. Notwithstanding globalization, the US model of contemporary capitalism is not a global model. No other contemporary capitalist economy has made the commitment to 'shareholder value' that is the most distinctive feature of the US model. Japan has come through the stagnation of the 1990s as a highly innovative economy, while eschewing shareholder value ideology and practices (Lazonick 2005). In western European nations the ideology has been tempered by a commitment to 'social inclusion'; the question is whether the equity and stability that social inclusion brings can be harnessed to support innovative enterprise. In the emerging giants, India and China, the stock market has come to play a more important, and possibly dangerous, role. In all of these economies, the success of innovative companies has been based, however, not on the stock market, but on the principles of strategic control, organizational integration, and financial commitment. Historically these principles also underpinned innovative enterprise in the United States. Many corporate executives who exercise control over resource allocation in the US economy may, however, have forgotten these principles, or worse yet, while they have been busy enriching themselves, they may have never bothered to learn them.

Bibliography

To conserve both the word-count and flow of this essay, I have kept bibliographic references to a minimum, indicating instead works of mine in which these references can be found.

Carpenter, M., W. Lazonick, and M. O'sullivan. 2003. The stock market and innovative capability in the New Economy: The optical networking industry. *Industrial and Corporate Change* 12: 963–1034.
Council of Economic Advisers. 2006. *Economic report of the president, 2006*. Washington, DC: Executive Office of the President.
Electronic News. 2005. US could lose race for nanotech leadership, SIA panel says. 16 March. Online. Available at http://www.reed-electronics.com/electronicnews/article/CA511197?nid=2019&rid = 1344283927. Accessed 5 Sept 2006.
Lazonick, W. 2003. The theory of the market economy and the social foundations of innovative enterprise. *Economic and Industrial Democracy* 24: 9–44.
Lazonick, W. 2004. Corporate restructuring. In *The Oxford handbook of work and organization*, ed. S. Ackroyd et al. Oxford: Oxford University Press.
Lazonick, W. 2005. The institutional triad and Japanese development [translated into Japanese]. In *The contemporary Japanese enterprise*, vol. 1, ed. G. Hook and A. Kudo. Tokyo: Yukikaku Publishing.
Lazonick, W. 2006a. Evolution of the new economy business model. In *Internet and digital economics*, ed. E. Brousseau and N. Curien. Cambridge: Cambridge University Press.
Lazonick, W. 2006b. Globalization of the ICT labor force. In *The Oxford handbook on ICTs*, ed. R. Mansell et al. Oxford: Oxford University Press.
Lazonick, W. 2006c. Innovative enterprise and economic development. In *Business performance in twentieth century: a comparative perspective*, ed. Y. Cassis and A. Colli. Cambridge: Cambridge University Press.
Lazonick, W. 2006d. *The US stock market and the governance of innovative enterprise*. INSEAD: Working paper.
Lazonick, W., and M. O'sullivan. 2000. Maximizing shareholder value: a new ideology for corporate governance. *Economy and Society* 29: 13–35.
New York Times. 1996. *The downsizing of America*. New York: Times Books.
O'sullivan, M. 2000. The innovative enterprise and corporate governance. *Cambridge Journal of Economics* 24: 393–416.
Pisano, G. 2006. *The science bsiness: Strategy, organization, and leadership in biotechnology*. Cambridge, MA: Harvard Business School Press.
Saxenian, A. 2006. *The new argonauts: Regional advantage in a global economy*. Cambridge, MA: Harvard University Press.
Schumpeter, J. 1934. *The theory of economic development*. Oxford: Oxford University Press.
Schumpeter, J. 1950. *Capitalism, socialism and democracy*. 3rd ed. New York: Harper.
Wbur.org. 2006. Sharpening the cutting edge. *On Point*. 9 February. Online. Available at http://www.onpointradio.org/shows/2006/02/20060209_b_main.asp. Accessed 5 Sept 2006.
Whyte, W. 1956. *The organization man*. New York: Simon and Schuster.

Contestable Markets

Robert D. Willig

C

Keywords

Antitrust enforcement; Barriers to entry; Contestable markets; Increasing returns; Integer problem; Long-run competitive equilibrium; Marginal and average cost pricing; Natural monopoly; Oligopoly; Partial equilibrium; Perfect competition; Perfectly contestable markets; Ramsey optimum; Scale economies; Weak invisible hand theorem of natural monopoly

JEL Classifications

D4

Contestable markets are those in which competitive pressures from potential entrants exercise strong constraints on the behaviour of incumbent suppliers. For a market to be contestable, there must be no significant entry barriers. Then, in order to offer no profitable opportunities for additional entry, an equilibrium configuration of the industry must entail no significant excess profits, and must be efficient in its pricing and in its allocation of production among incumbent suppliers. This is so of a contestable market whether it is populated with only a monopolist or with a large number of actively competing firms, because it is potential competition from potential entrants rather than competition among active suppliers that effectively constrains the equilibrium behaviour of the incumbents.

Perfectly contestable markets (PCMs) are a benchmark for the analysis of industry structure – a benchmark based on an idealized limiting case. Perfectly contestable markets are open to entry by entrepreneurs who face no disadvantages vis-à-vis incumbent firms and who can exit without loss of any costs that entry required to be sunk. The potential entrants have available the same

best-practice production technology, the same input markets and the same input prices as those available to the incumbents. There are no legal restrictions on market entry and exit, and there are no special costs that must be borne by an entrant that do not fall on incumbent firms as well. Consumers have no preferences among firms except those arising directly from price or quality differences in firms' offerings.

Potential entrants into perfectly contestable markets are profit-seekers who respond with production to profitable opportunities for entry. They assess the profitability of their marketing plans by making use of the current prices of incumbent firms. Thus, for example, an entrepreneur will enter a market if he anticipates positive profit from undercutting the incumbent's price and serving the entire market demand at the new lower price. Potential entrants are undeterred by prospects of retaliatory price cuts by incumbents and, instead, are deterred only when the existing market prices leave them no room for profitable entry.

These features of the behaviour of potential entrants are key to the workings of perfectly contestable markets, and they are fully rational only where entry faces no disadvantages and is costlessly reversible. Hence, the benchmark case of perfect contestability excludes the sunk costs, precommitments, asymmetric information and strategic behaviour that characterize many real markets and that are the focus of much current research attention in the field of industrial organization. With irreversibilities and the inducements for strategic behaviour absent, industry structure in PCMs is determined by the fundamental forces of demand and production technology.

Of course, this is also true of perfectly competitive markets. However, this most familiar idealized limiting case is not a satisfactory benchmark for the study of industry structure in general, because it is intrinsically inapplicable to a variety of significant cases. In particular, where increasing returns to scale are present, perfectly competitive behaviour is logically inconsistent with the long-run financial viability of unsubsidized firms.

Perfectly contestable markets can serve in place of perfectly competitive markets as the general standard of comparison for the organization of industry whether or not scale economies are prevalent. Where they are not, perfectly competitive behaviour is necessary for equilibrium in PCMs, and, where scale economies do prevail, equilibrium in PCMs entails behaviour different from that found in perfectly competitive markets but which none the less tends to exhibit desirable welfare properties. In other words, perfect contestability is a generalization of perfect competition that has strong implications in significant circumstances where the latter is inapplicable.

In order to clarify and expand on these ideas, subsequent sections offer analytic outlines of the theory of perfectly contestable markets and applications of the theory to single-product and multiproduct industries. Finally, observations are offered on the implications of this theory for the formulation of government policy towards industry.

Perfectly Contestable Markets: Definitions and Basic Properties

The theory presented here lies in the realm of partial equilibrium. It deals with the provision of the set of products $N = \{1, \ldots, n\}$, some of which may not actually be produced, and which is a proper subset of all the goods in the economy. The prices of these products are represented by vectors $p \in \mathrm{R}^n_{++}$, and other prices are assumed to be exogenous and are suppressed in the notation. $Q(p) \in R^n_+$ is the vector-valued market demand function for the products in N, and it suppresses consumers' incomes which are assumed to be exogenous. For any output vector $y \in \mathrm{R}^n_+$, $C(y)$ is the cost at exogenously fixed factor prices when production is efficient. The underlying technology is assumed to be freely available to all incumbent firms and all potential entrants. Where necessary, $C(y)$ and $Q(p)$ will be assumed to be differentiable.

Definition 1 A *feasible industry configuration* is composed of m firms producing output vectors $y^1, \ldots, y^m \in \mathrm{R}^n_+$, at prices $p \in R^n_{++}$ such that the markets clear, $\sum_{i=1}^m y^i = Q(p)$, and that each firm at least breaks even, $p \cdot y^i - C(y^i) \geq 0, i = 1, \ldots, m$.

Thus, the industry configuration is taken to be comprised of m firms, where m can be any positive integer, so that the industry structure is monopolistic if $m = 1$, competitive if m is sufficiently large, or oligopolistic for intermediate values of m. The term 'feasibility' refers to the requirements that each of the firms involved selects a non-negative output vector that permits its production costs, $C(y^i)$ to be covered at the market prices, p, and that the sum of the outputs of the m firms satisfies market demands at those prices.

Definition 2 A feasible industry configuration over N, with prices p and firms' outputs $y^1,\ldots,y^m$, is *sustainable* if $p^e \,.\, y^e \leq C(y^e)$, for all $p^e \in R^n_{++}$, $y^e \in R^n_+$, $p^e \leq p$, and $y^e \leq Q(p^e)$.

The interpretation of this definition is that a sustainable configuration affords no profitable opportunities for entry by potential entrants who regard incumbents' prices as fixed (for a period sufficiently long to make $C(\cdot)$ the relevant flow cost function for an entrant). Here, a feasible marketing plan of a potential entrant is comprised of prices, p^e, that do not exceed the incumbents' quoted prices, p, and a quantity vector, y^e, that does not exceed market demand at the entrant's prices, $Q(p^e)$. The configuration is sustainable if no such marketing plan for an entrant offers a flow of profit, $p^e \,.\, y^e - C(y^e)$, that is positive.

Definition 3 A *perfectly contestable market* (PCM) is one in which a necessary condition for an industry configuration to be in equilibrium is that it be sustainable.

A PCM so defined may be interpreted, heuristically, as a market subject to potential entry by firms that have no disadvantage relative to incumbents, and that assess the profitability of entry on the supposition that incumbents' prices are fixed for a sufficiently long period of time. Then, since one requirement for equilibrium is the absence of new entry, an equilibrium configuration in a PCM must offer no inducement for entry; that is, it must be sustainable.

Definition 4 A feasible industry configuration over N, p; $y^1,\ldots, y^m$, is a *long-run competitive equilibrium* if $p \,.\, y \leq C(y)\ \forall\ y \in R^n_+$.

So defined, a long-run competitive equilibrium has precisely the characteristics usually ascribed to it. Together, $p \,.\, y^i \geq C(y^i)$ and $p.y \leq C(y), \forall y \in R^n_+$, imply that $p \,.\, y^i = C(y^i)$ and that the $y^i \in \arg\max_y [p \,.\, y - C(y)]$. Thus, each firm in the configuration takes prices as parametric, chooses output to maximize profits, earns zero profit, and equates marginal costs to prices of produced outputs. It is now easy to show

Proposition 1 A long-run competitive equilibrium is a sustainable configuration, so that a perfectly competitive market is a PCM.

Proposition 2 Sustainable configurations need not be long-run competitive equilibria, and a PCM need not be perfectly competitive.

The simplest example sufficient to prove this second proposition is an industry producing a single product with increasing returns to scale over the relevant range of output. Here, the only feasible configuration that is sustainable entails one firm producing the maximal output level y^* given by the intersection of the declining average cost curve with the industry demand curve, and selling at the price p^* given by the corresponding level of average cost. This configuration is sustainable because, at a price equal to or less than p^*, sale of any quantity on or inside the demand curve yields revenue no greater than production cost; in this range, price does not exceed average cost. Yet this configuration is not a long-run competitive equilibrium, as defined above, because sale of a quantity greater than y^* would earn positive profit if the price could remain at p^*, and because at y^* price exceeds marginal cost which is less than average cost. In fact, in this example there is no possible long-run competitive equilibrium since marginal cost lies below average cost throughout the relevant range of output levels given by demand. In contrast, there is a sustainable configuration.

Hence, Propositions 1 and 2 show that the sustainable industry configuration is a substantive generalization of the long-run competitive equilibrium, and that the PCM is a substantive generalization of the perfectly competitive market. The

following propositions summarize some characteristics of equilibria in PCMs.

Proposition 3 Let $p; y^1,\ldots, y^m$ be a sustainable industry configuration. Then each firm must (i) earn zero profit by operating efficiently, $p \cdot y^i - C(y^i) = 0$; (ii) avoid cross-subsidization, $p_s \cdot y_s^i \geq C(y^i) - C\left(y_{N-s}^i\right), \forall SCN$ (where the vector x_T agrees with the vector x in components $j \in T$ and has zeros for its other components); (iii) price at or above marginal cost, $p_j \geq \partial C(y^i)/\partial y_j$.

The interpretation of condition (ii) is that the revenues earned from the sales of any subset of the goods must not fall short of the incremental costs of producing that subset. Otherwise, in view of the equality of total revenues and costs, the revenues collected from the sales of the other goods must exceed their total stand-alone production cost. In PCMs, such pricing invites entry into the markets for the goods providing the subsidy.

Proposition 4 Let $p; y^1,\ldots, y^m$ be a sustainable configuration with $y_j^k < \sum_{h=1}^m y_j^h$. Then $p_j = \partial C(y^k)/\partial y_j$. That is, if two or more firms produce a given good in a PCM, they must select input–output vectors at which their marginal costs of producing it are equal to the good's market price.

The implications of this result are surprisingly strong. The discipline of sustainability in perfectly contestable markets forces firms to adopt prices just equal to marginal costs, provided only that they are not monopolists of the products in qst. Conventional wisdom implies that, generally, only perfect competition involving a multitude of firms, each small in its output markets, can be relied upon to provide marginal-cost prices. Here we see that potential competition by prospective entrants, rather than rivalry among incumbent firms, suffices to make marginal-cost pricing a requirement of equilibrium in PCMs, even those containing as few as two active producers of each product. The conventional view holds that the enforcement mechanism of full competitive equilibrium requires the smallness of each active firm in its product market, in addition to freedom of entry. We see that the smallness requirement can be dispensed with, almost entirely, with exclusive reliance on the freedom of entry that characterizes PCMs.

Proposition 5 Let $p; y^1,\ldots, y^m$ be a sustainable configuration. Then, for any $\hat{y}^1, \ldots, \hat{y}^k$ with

$$\sum_{j=1}^{k} \hat{y}^j = \sum_{j=1}^{m} y^j, \sum_{j=1}^{k} C\left(\hat{y}^j\right) \geq \sum_{j=1}^{m} C(y^j).$$

That is, a sustainable configuration minimizes the total cost to the industry of producing the total industry output.

This proposition is a generalization to PCMs of a well-known result for perfect competition. It can be interpreted as a manifestation of the power of unimpeded potential entry to impose efficiency upon the industry. For example, the proposition implies that if a monopoly occupies a PCM it must be a *natural* monopoly – production by a single firm must minimize industry cost for the given output vector. Thus, Propositions 3, 4 and 5 are powerful tools for the analysis of industry structure in PCMs. Proposition 5 permits information on the properties of production costs to be used to assess the scale and scope of firms' activities in PCMs. Then, Propositions 3 and 4 permit inferences to be drawn about the corresponding equilibrium prices.

PCMS with a Single Product

This analytic approach leads to very strong results in the single-product case. Propositions 3–5 show that there are only two possible types of sustainable configurations in single-product industries. The first type involves a single firm which charges the lowest price that is consistent with nonnegative profit. The firm must be a natural monopoly when it produces the quantity that is demanded at this price. And, in this circumstance, the result maximizes welfare subject to the constraint that all firms in question be viable financially without subsidies. Such a second-best maximum is referred to as a 'Ramsey optimum'.

The second type of sustainable configuration involves production by one or more firms of

outputs at which both marginal cost and average cost are equal to price. Here, in the long run, all active firms exhibit the behaviour that characterizes perfectly competitive equilibrium. And, of course, the result involves both (first-best) welfare optimality and financial viability. Hence, in this case, Ramsey optimality and the first-best coincide. This establishes the result that in a single-product industry any sustainable configuration is Ramsey optimal.

However, in general, because of the 'integer problem', sustainable configurations may generally not exist. This problem arises, for example, where there is only one output at which a firm's marginal and average costs coincide, and where the quantity of output demanded by the market at the competitive price is greater than this, but is not an integer multiple of that amount. Then, no sustainable configurations exist.

There is, however, a plausible assumption, supported by empirical evidence, at least to some degree, that eliminates the integer problem. Suppose that a firm's average cost curve has a flat-bottom rather than being 'U'-shaped. In particular, suppose that the minimum level of average cost is attained not only at one output, but (at least) at all outputs between the minimum efficient scale, y_m, and twice the minimum efficient scale. Then any industry output, y^1, that is at least equal to y_m can be apportioned among an integer number of firms, each of which achieves minimum average cost. Specifically, y^1 can be divided evenly among $\lfloor y^1/y^m \rfloor$ firms (where $\lfloor x \rfloor$ is the largest integer not greater than x) and each firm's output, $y^1/\lfloor y^1/y^m \rfloor$ must lie in the (half-open) interval between y_m and $2y_m$. Hence, in this case, the Ramsey optimum can either be a sustainable configuration of two or more firms performing competitively, or a sustainable natural monopoly. Such a monopoly may either produce an output at which there are increasing returns to scale and it will then price at average cost, or it may produce an output between y_m and $2y_m$ with locally constant returns to scale and adopt a price equal both to average and marginal cost. This, together with the preceding argument, establishes the following result.

Proposition 6 In a single-product industry in which the firm's average cost curve has a flat-bottom between minimum efficient scale and twice minimum efficient scale, a configuration is sustainable if and only if it is Ramsey optimal.

This result shows that, under the conditions described, there is equivalence between welfare optimality and equilibrium in PCMs. This extends the corresponding result for perfectly competitive equilibria to cases of increasing returns to scale. Moreover, since the behavioural assumptions required for a PCM are weaker than those underlying perfectly competitive markets, the equivalence result is more sweeping. In particular, Proposition 6 implies that PCMs can be expected to perform well, whatever the number of firms participating in equilibrium. It is the potential competition of potential entrants, rather than the active competition of existing rivals, that drives equilibrium in PCMs with a single product to welfare optimality.

Multi-Product Perfectly Contestable Markets

In industries that produce two or more goods, a rich variety of industry structures become possible, even in PCMs. Here, while the constraints imposed upon incumbents by perfect contestability are not nearly as effective in limiting the range of possible outcomes as they are in single product industries, they nevertheless provide a helpful basis for analysis. In particular, Propositions 3–5 indicate connections among various qualitative properties of multi-product cost functions and various elements of industry structure in PCMs. These connections constitute one theme of this section. The other theme is the normative evaluation of the industry structures that arise in multi-product PCMs.

Before proceeding, it may be useful to provide definitions of some of the multiproduct cost properties that are used in the analysis.

Definition 5 Let $P = \{T_1, \ldots, T_k\}$ be a non-trivial partition of $S \subseteq N$. There are (weak)

economies of scope at y_s with respect to the partition P if $\sum_{i=1}^{k} C(y_{Ti}) > (\geq) C(y_s)$. If no partition is mentioned explicitly, then it is presumed that $T_i = \{i\}$.

Definition 6 The *degree of scale economies* defined over the entire product set, $N = \{1,\ldots, n\}$, at y, is given by $S_N(y) = C(y)/y.\nabla C(y)$.

Returns to scale are said to be increasing, constant or decreasing as S_N is greater than, equal to or less than unity. This occurs as the elasticity of ray average cost with respect to t is negative, positive or zero; where *ray average cost* is RAC $(ty^0) \equiv C(ty^0)/t$.

Definition 7 The *incremental cost* of the product set $T \subseteq N$ at y is given by $IC_T(y) \equiv C(y) - C(y_{N-T})$. The *average incremental cost* of T is $AIC_T(y) \equiv IC_T(y)/\sum_{j \in T} y_j$.

The average incremental cost of T is decreasing, increasing or constant at y if $AIC_T(ty_T + y_{N-T})$ is a decreasing, increasing or locally constant function of t at $t = 1$. These cases are labelled respectively, increasing, decreasing or constant *returns to the scale of the product line T*. The degree of scale economies specific to T is

$$IC_T(y) / \sum_{i \in T} y_i \frac{\partial C(y)}{\partial y_i}.$$

Definition 8 A cost function $C(y)$ is *trans-ray convex* through some point $y^* = (y_1^*, \ldots, y_n^*)$ if there exists at least one vector of positive constants $w_1, \ldots, w_n$ such that for every two output vectors $y^a = (y_1^a, \ldots y_n^a)$ and $y^b = (y_1^b, \ldots y_n^b)$ that lie on the hyperplane $\sum w_i y_i = w_0$ through point y^*, $C[ky^a + (1-k)y^b] \leq kC(y^a) + (1-k)C(y^b)$ for $k \in (0, 1)$.

In view of the general result that sustainable configurations minimize industry-wide costs (Proposition 5), these cost properties permit inferences to be drawn about industry structure in multi-product PCMs. The first issue that arises is when multicommodity production is characteristic of equilibrium in a PCM.

Proposition 7 A multi-product firm in a PCM must enjoy (at least weak) economies of scope over the set of goods it produces. When strict economies of scope are present, there must be at least one multi-product firm in any PCM that supplies more than one good.

The second basic question that arises is whether there can be two or more firms actively producing a particular good in a PCM. If there are, then, by Proposition 4, marginal cost pricing must result. The answer depends upon the availability of product-specific scale economies.

Proposition 8 Any product with average incremental costs that decline throughout the relevant range (that is, that offers product-specific increasing returns to scale) must be produced by only a single firm (if it is produced at all) in a PCM. Further, such a product must be priced above marginal cost, unless the degree of product- specific scale economies is exactly one.

Thus, regardless of the presence or absence of economies of scope, globally declining average incremental costs imply that a product must be monopolized in a PCM. It is an immediate corollary that if all goods in the set N exhibit product- specific scale economies, and if there are economies of scope among them all, then the industry is a natural monopoly that must be monopolized in a PCM.

Another route to this result is provided by the 'weak invisible hand theorem of natural monopoly'.

Proposition 9 Trans-ray convexity of costs together with global economies of scale imply natural monopoly. If, in addition certain other technical conditions are met, a monopoly charging Ramsey-optimal prices is a sustainable configuration.

In general, there may exist natural monopoly situations in which no sustainable prices are possible for the Ramsey optimal product set. Further, even where sustainable prices exist, the Ramsey optimal prices may not be among them. However, under the conditions of the weak invisible hand theorem, the Ramsey optimal prices for the Ramsey optimal product set are guaranteed to be sustainable, so that PCMs are consistent with (second-best) welfare optimal performance by a natural monopoly.

PCMs will yield first-best welfare optimality if there exist sustainable configurations with at least

two firms actively producing each good. For in this case Propositions 4 and 5 guarantee industry-wide cost efficiency and marginal-cost pricing of all products. Here, two issues must be resolved: Does industry-wide cost minimization require at least two producers of each good? And if so, do sustainable configurations exist?

The existence problem can be solved in a manner analogous to its solution in the case of single-product industries: by assuming that ray average costs remain at their minimum levels for output vectors that lie (on each ray) between minimum efficient scale and twice minimum efficient scale. And the presence of at least two producers (or one operating in the region where constant returns prevail) of each good is assured if the quantities demanded by the market at the relevant marginal-cost prices are no smaller than minimum efficient scale (along the relevant ray) and if the cost function exhibits trans-ray convexity.

Policy Implications of PCMS

One of the principal lessons of the analysis of PCMs is that monopoly does not necessarily entail welfare losses. Rather, the 'weak invisible hand th' shows that under certain conditions sustainability and Ramsey optimality are consistent, so that the total of consumers' and producers' surpluses may well be maximized (subject to the constraint that firms be self-supporting) in the equilibrium of a monopoly which operates in contestable markets.

Even stronger results follow from the discussed results that under certain conditions sustainability and a first-best solution are consistent in an oligopoly with a small number of firms. When minimization of industry cost requires that each good be produced by at least two firms, sustainability requires any equilibrium to satisfy the necessary conditions for a first-best allocation of resources. Thus, in these cases, the invisible hand has the same power over oligopoly in perfectly contestable markets that it exercises over a perfectly competitive industry.

This theory suggests that in a market that approximates perfect contestability, the general public interest is well-served by a policy of laissez-faire rather than active regulation by administrative or antitrust means. Small numbers of large firms, vertical and even horizontal mergers and other arrangements which have traditionally been objects of suspicion of monopolistic power, are rendered harmless and perhaps even beneficial by the presence of contestability.

On the other hand, contestability theory does not lend support to the proposition that the unrestrained market automatically solves all economic problems and that virtually all regulation and antitrust activity entails unwarranted and costly intervention. The economy of reality is composed of industries which vary widely in the degree to which they approximate the attributes of perfect contestability. Before the theory of contestability can be legitimately applied to reach a conclusion that intervention is unwarranted in a specific sector, it must first be shown that the sector lies unprotected by entry barriers and that the force of potential entry therefore actively constrains the behaviour of incumbent firms. This then becomes the appropriate first stage in an analysis of efficient government policy towards an industry. Only where the conditions of contestability are found to characterize the reality of an industry can there be validity in applying the normative conclusions of contestability theory concerning the power of potential competition actually to enforce efficient behaviour on incumbents.

Even where contestability is absent in reality, the formulation on efficient regulation can be usefully guided by the theory of contestable markets instead of the theory of perfectly competitive markets. The first-best lesson of the perfect competition model, calling for prices to be set equal to marginal costs, has no doubt contributed to the common regulatory ethos which *equates* price to *some* measure of cost. This doctrine has been used frequently where it is completely inappropriate and without logical foundation, that is, in cases where prices should be based on demand as well as cost considerations, because of the presence of economies of scale and scope. Such arbitrary measures as fully distributed costs cannot substitute for marginal cost measures as decision rules for proper pricing, and the search for a substitute is a remnant of inappropriate reliance on the

model of perfect competition for guidance in regulation.

In contrast, contestability theory suggests cost measures that are appropriate guideposts for regulated pricing – incremental and stand-alone costs. The incremental cost of a given service is, of course, the increment in the total costs of the supplying firm when that service is added to its product line. In perfectly contestable markets, the price of a product will lie somewhere between its incremental and its stand-alone cost, just where it falls in that range depending on the state of demand. One cannot legitimately infer that monopoly power is exercised from data showing that prices do not exceed stand-alone costs, and stand-alone costs constitute the proper cost-based ceilings upon prices, preventing both cross-subsidization and the exercise of monopoly power. A simple example will show why this is so.

First, suppose that a firm supplies two services, A and B, which *share no costs* and that each costs 10 units a year to supply. The availability of effective potential competition would force revenues from each service to equal 10 units a year. For higher earnings would attract (profitable) entry, and lower revenues would drive the supplier out of business. In this case, in which common costs are absent, incremental and stand-alone costs are equal to each other and to revenues, and the competitive and contestable benchmarks yield the same results.

Next, suppose instead that of the 20 unit total cost 4 are fixed and common to A and B, while 16 are variable, 8 of the 16 being attributable to A and 8 to B. If, because of demand conditions, at most only a bit more than 8 can be generated from consumers of A, then a firm operating and surviving in contestable markets will earn a bit less than 12 from B. These prices lie between incremental costs (8) and standalone costs (12), are mutually advantageous to consumers of both services, and will attract no entrants, even in the absence of any entry barriers. In contrast, should the firm attempt to raise the revenues obtained from B above the 12 unit stand-alone cost, it would lose its business to competitors willing to charge less. Similarly, the same fate would befall it in contestable markets if it priced B in a way that earned more than 8 plus the common cost of 4, less the contribution towards that common cost from service A.

Thus, the forces of idealized potential competition in perfectly contestable markets enforce cost constraints on prices, but prices remain sensitive to demands as well. Actual competition and potential competition are *effective* if they constrain rates in this way, and in such circumstances regulatory intervention is completely unwarranted. But if, in fact, market forces are not sufficiently strong, then there may be a proper role for regulation of natural monopoly, and the theoretical guidelines derived from the workings of contestable markets are the appropriate ones to apply. That is, prices must be constrained to lie between incremental and stand-alone costs. (This is the approach recently adopted by the Interstate Commerce Commission to determine maximum rates for US railroad services, and the method has already withstood appeals to the federal courts.)

See Also

► Barriers to Entry

Bibliography

Baumol, W.J., and R.D. Willig. 1986. Contestability: Developments since the book. *Oxford Economic Papers* 38: 9–36.

Baumol, W.J., J.C. Panzar, and R.D. Willig. 1982. *Contestable markets and the theory of industry structure*. New York: Harcourt Brace Jovanovich.

Baumol, W.J., J.C. Panzar, and R.D. Willig. 1985. On the theory of perfectly contestable markets. In *New developments in the analysis of market structure*, ed. J. Stiglitz and F. Mathewson. New York: Harcourt Brace.

Contingent Commodities

Zvi Safra

Keywords

Contingent commodities; Convexity; General equilibrium; Risk; Risk aversion; Uncertainty

JEL Classification
D5

The theory of general competitive equilibrium was originally developed for environments where no uncertainty prevailed. Everything was certain and phrases like 'it might rain' or 'the weather might be hot' were outside the scope of the theory. The idea of *contingent commodity,* that was introduced by Arrow (1953) and further developed by Debreu (1953), was an ingenious device that enabled the theory to be interpreted to cover the case of uncertainty about the availability of resources and about consumption and production possibilities. Basically, the idea of contingent commodity is to add the environmental event in which the commodity is made available to the other specifications of the commodity. With no uncertainty every commodity is specified by its physical characteristics and by the location and date of its availability. It is fairly clear, however, that such a commodity can be considered to be quite different where two different environmental events have been realized. The following examples clarify this: an umbrella at a particular location and at a given date in case of rain is clearly different from the same umbrella at the same location and date when there is no rain; some ice cream when the weather is hot is clearly different from the same ice cream (and at the same location and date) when the weather is cold; finally, the economic role of wheat with specified physical characteristics available at some location and date clearly depends on the precipitation during its growing season. Thus, specifying commodities by both the standard characteristics and the environmental events seems very natural, whereas the role of the adjective in 'contingent commodities' is simply to make it clear that one is dealing with commodities the availability of which is contingent on the occurrence of some environmental event. With this specification the model with contingent commodities is very similar to the classical model of general competitive equilibrium and thus questions like the existence of equilibrium and its optimality (with the additional aspect of efficient allocation of risk bearing) are answered in a similar way. Note that, although this model deals with uncertainty, no concept of probabilities is needed for its formal description.

To make things more explicit we look at a simple model with contingent commodities. Assume that, without referring to uncertain events, there are $k \geq 1$ commodities, indexed by i, and that there are $n > 1$ mutually exclusive and jointly exhaustive events (or states of nature), indexed by s, where k and n are finite. Thus a contingent commodity is denoted by x_{is} and the total number of these commodities is kn, which is greater than k but still finite. Consumption and production sets are thus defined as subsets of the kn-dimensional Euclidean space, and the economic behaviour of firms and consumers naturally follows from profit maximization (by firms) and utility maximization (by consumers). The price p_{is} of the contingent commodity x_{is} is the number of units of account that have to be paid in order to have the ith commodity being delivered at the sth event. It is assumed that the market is organized before the realization of the possible events. Thus payment for the contingent commodity x_{is} is done at the beginning while delivery takes place after the realization of events and only in case event s has occurred. Note that the price of the (certain) ith commodity, that is, the number of units of account that have to be paid in order to have the ith commodity *for sure,* is the sum over s of the prices p_{is}. For example, assume that the price of one quart of ice cream if the weather is hot is \$2.00, the price of one quart of the same ice cream if the weather is cold is \$1.00 and that $n = 2$ (either it is hot or cold). Thus the price of having one quart of that ice cream for sure is \$2.00 + \$1.00 = \$3.00.

It should be noted that, although the probabilities of the possible events do not explicitly enter the model, the attitude towards risk of both consumers and producers is of interest and does play a significant role in this framework. The preference relations of consumers defined on subsets of the kn-dimensional Euclidean space reflect not only their 'tastes' but also their subjective beliefs about the likelihoods of different events as well as their attitude towards risk. Convexity of consumers' preferences, for example, is interpreted as risk aversion while, in the same spirit, profit maximization

of firms is interpreted as risk neutrality. It should be mentioned that both Arrow and Debreu basically assume expected utility maximizing behaviour, in the sense of the Savage (1954) framework. A more general approach to such preference relations can be found in Yaari (1969), where, again, convexity is taken to mean risk aversion.

A unified and more formal treatment of time and uncertainty using contingent commodities can be found in Debreu (1959, ch. 7). Radner (1968) presents an extension of the above model to the case in which different economic agents have different information.

See Also

- ▶ Arrow–Debreu Model of General Equilibrium
- ▶ Uncertainty
- ▶ Uncertainty and General Equilibrium

Bibliography

Arrow, K.J. 1953. Le rôle de valeurs boursières pour la répartition la meilleure des risques. *Econométrie.* Paris: CNRS. English translation 'The role of securities in the optimal allocation of risk-bearing' in *Review of Economic Studies* (1964); reprinted in K.J. Arrow, *Essays in the theory of risk-bearing.* Chicago: Markham, 1971.

Debreu, G. 1953. Une économie de l'incertain. Mimeo, Paris: Electricité de France.

Debreu, G. 1959. *Theory of value*. New York: Wiley.

Radner, R. 1968. Competitive equilibrium under uncertainty. *Econometrica* 36: 31–58.

Savage, L.J. 1954. *Foundations of statistics*. New York: Wiley.

Yaari, M.E. 1969. Some remarks on measures of risk aversion and their uses. *Journal of Economic Theory* 1: 315–329.

Contingent Valuation

Trudy Ann Cameron

Abstract

'Contingent valuation' methods are used to generate demand data, usually from household surveys, when real markets do not supply reliable revealed preference data about demands for certain types of goods. A number of significant lawsuits have promoted their use in estimating demand for environmental goods. They are also used by transportation economists, health economists and market researchers. Although the degree of acceptance of these methods varies, many economists agree that a value based on stated preferences derived from carefully conceived and executed research is almost certainly preferable to no number at all.

Keywords

Conjoint analysis; Construct validity assessments; Contingent valuation; Discrete- choice models; Dummy variables; Environmental economics; Expectations; Health economics; Household surveys; Indirect utility functions; Logit models; Marginal utility of income; Maximum likelihood; Neuroeconomics; Probit models; Random utility models; Revealed preference; Scope tests; Stated preference; Willingness to pay

JEL Classifications

Q51

Most economists would agree that no researcher should prefer demand data from hypothetical markets if data concerning the identical goods or services, based on real markets, are readily available. However, there are many situations when even the cleverest empirical economist cannot come up with revealed preference data from actual markets that can be relied upon for information about household demands for some types of goods. Environmental goods are one class of goods where real-market demands sometimes cannot be measured adequately. In the 1980s, environmental economists began in earnest to exploit stated-preference demand information, usually collected using household surveys. This demand information is used primarily to produce utility-theoretic measures of the social benefits of environmental protection measures for benefit–cost analyses.

Environmental economists called these methods 'contingent valuation methods' (CVM) because the valuations were elicited 'contingent upon the conditions described in the survey'. Research that focused on the development and assessment of CVM in environmental economics was well under way by the mid-1980s. However, two events in 1989 thrust the method to the forefront of the field. First, the *Exxon Valdez,* an ocean tanker, ran aground in Prince William Sound in Alaska, spilling 11 million gallons of oil in an environmental disaster that attracted a huge amount of media attention worldwide. Second, just a few months later, the US Court of Appeals held that the economic damages assessed for spills of oil or other hazardous substances could include 'lost passive use values', and that these values could legally be measured via CVM.

Plaintiffs and defendants in the Exxon Valdez case thus had a big incentive to advocate and derogate CVM, respectively. For at least a time, the discussion in the literature teetered on the brink of losing its polite academic tone. Given the escalation of the controversy over CVM, the US National Oceanic and Atmospheric Administration (NOAA) convened a panel of experts (untainted by any active role in the Exxon Valdez litigation) to assess CVM. This exercise, by Arrow et al. (1993), produced a set of pronouncements concerning best practices for the conduct of CVM studies. While the 1993 NOAA Panel report cannot be considered the last word on CVM, it was very influential, and there has since been strong pressure on researchers either to conform to the NOAA best practices or to fully justify any departures from them.

As a result of the Exxon Valdez case, much doubt about the reliability of stated preference data led to numerous comparisons of the implications of stated and revealed preference data (for example, Carson et al. 1996). CVM works best when respondents have a clear sense of the consequences of their choices – in terms of both their own budgets and the exact nature of the good that they are being asked to consider paying for – and when they are reasonably familiar with market transactions involving that good. This means that CVM is, unfortunately, most successful when it is least needed. The challenge for researchers is to ensure that demand information gathered using CVM, in less-than-ideal contexts, is as valid and useful as possible.

Myriad biases and qualifications may afflict poorly executed CVM studies. A partial list includes incentive compatibility, hypothetical bias (if the choice is perceived to have absolutely no real consequences), strategic bias (when people try to manipulate the outcome by misrepresenting their preferences), non-response bias (since people cannot be compelled to participate), starting-point bias (for surveys with iterative bids), interviewer bias (for in-person surveys), and information bias (when some portion of the value is constructed during the survey where it did not exist before). Other problems include yea-saying, part–whole bias or embedding, scenario rejection, and the potential for respondents not to pay sufficient attention to their real budget constraints.

Choice formats have been an important issue in the development of CVM. For example, in some early applications of CVM survey respondents were asked directly to identify the single highest dollar amount that they would pay to obtain some change in conditions. These were called open-ended CV questions. Researchers quickly realized that such a task was difficult for consumers who were unfamiliar with naming their own price, especially for goods they may never before have thought much about having to pay for. CVM elicitation techniques evolved fairly quickly to a dichotomous-choice format, where respondents are given a choice between two states of the world. One state is typically the status quo, while the other involves a specified change or set of changes (such as an improvement in environmental quality, or some other rationed public good) that come at a price (typically a lump-sum payment). This binary choice format was found to fit naturally into a random utility model (RUM) framework that had also become a popular approach to consumer choice problems, both real and hypothetical, in the transportation mode-choice literature and elsewhere in economics.

Respondents' preferences, based on their answers to dichotomous-choice CV questions,

can be characterized either in an ad hoc fashion or in a more formal utility-theoretic framework. One standard approach is to specify an indirect utility function shared by all respondents. In its simplest form, the level of indirect utility is assumed to depend on the individual's net income under each of the two alternatives, and upon a discrete indicator of whether there is a change in the rationed public good, or no change, under each alternative. Respondents can choose the environmental improvement programme along with its associated cost (implying lower net income), or decline the environmental improvement programme in order to avoid the cost (preserving their net income). If a respondent prefers the programme with its associated cost, the researcher assumes that the respondent's utility level is higher under that alternative. Equivalently, this means that the *net* indirect utility associated with the programme alternative is positive.

A discrete-choice econometric model, typically involving a binary logit estimator, is used to estimate the sample average marginal utilities of (*a*) net income and (*b*) the discrete bundle of changes represented by the programme in question. It is of course possible to allow for heterogeneity across the sample in these marginal utilities. Most often, heterogeneity is introduced by allowing the otherwise scalar marginal utility associated with going from 'no programme' to 'programme' to become a systematically varying parameter. Of course, if the identical programme is offered to all respondents, it is not possible to allow this marginal utility to vary with attributes of the programme. However, it can easily be allowed to vary with characteristics of the respondent.

Less commonly, the marginal utility of income is also allowed to vary across respondents, either with the respondent's income (to allow for diminishing marginal utility of income) or with other respondent characteristics. However, there is a premium on simplicity for the marginal utility of income, stemming from the need to use the estimated marginal utility of income parameter (s) to recover demand information. For this reason, many researchers will, if it is justified by the data, prefer a choice model that is linear and additively separable in net income under each alternative.

Linearity and additive separability in income is convenient (when warranted) because the willingness to pay (WTP) function associated with the fitted model is given by the marginal rate of substitution (MRS) between the programme and income. This MRS is given by the ratio of the marginal utility of the programme to the marginal utility of income, producing a result that can be expressed in dollars per 'unit' of the programme, where the program indicator is either zero or 1 in the simple binary case. In the non-stochastic case, for a simple dichotomous choice CVM model, this is a single number – 'WTP for the program' – if the researcher has assumed homogeneous preferences throughout the sample.

Some extra empirical housekeeping is necessary when it is acknowledged that this point estimate is constructed as the ratio of two estimated quantities, each of which (due to the use of maximum likelihood estimation methods for the logit or probit model) is an asymptotically normally distributed random variable. In theoretical terms, the ratio of two normally distributed random variables has an undefined mean, because zero is a possible value for the denominator. As a practical matter, some researchers use simulation methods to build up a sampling distribution for the estimated WTP. It is possible to use packaged software to make a large number of random draws from the joint distribution of the logit or probit parameters (based on the estimated parameter point estimates and the parameter variance–covariance matrix). One can then build up a sampling distribution for the needed ratio. Other strategies for dealing with this inconvenience involve estimating the model in 'WTP-space' rather than 'utility-space' or employing the newer mixed logit (random-parameters logit models) and stipulating that the marginal utility of income parameter be distributed lognormal (since it should be strictly positive, on average), rather than normal, so that the potential divide-by-zero problem goes away.

Over the 1990s contingent valuation researchers in environmental economics gradually made better contact with their counterparts working in other literatures who were confronted

by similar problems where there is a lack of market data for products or public goods that need to be valued. In the transportation literature, researchers had grappled early on with the problem of forecasting demand for public transportation projects, or new types of vehicles, that did not yet exist. Researchers began to introduce hypothetical new transportation options which could be characterized in the same terms as existing options (in 'attribute space') but which had some attributes that lay well outside the set of existing options on some dimensions, or which involved attributes that were not relevant for existing options (such as travel range or recharge time for prospective electric vehicles). One key difference from contingent valuation was the practice of asking survey respondents to consider more than just 'the status quo versus a single alternative'. Furthermore, the alternatives were more richly specified. Instead of using simply a dummy variable to indicate whether the policy, programme or public good was present or absent, each alternative was characterized in terms of an array of attributes.

Similar problems were also being addressed in the marketing literature, particularly in the context of 'pre-test' marketing. Companies considering whether to develop and introduce new products needed to know in advance about the likely demand for these products, perhaps as a function of alternative possible product configurations. Market researchers developed a set of techniques they called 'conjoint analysis'. In the marketing literature, the specifications used for the choice models were initially very ad hoc. Little attention was paid to the interpretation of the estimated coefficients as marginal utilities, and simple linear and additively separable specifications were very common. The slope coefficients were known as 'part-worths' rather than marginal utilities. However, much was learned about the degree of consistency between planned purchase behaviour and actual purchase behaviour.

CVM has also recently grown in popularity in other sub-disciplines, notably health economics. However, Smith (2003) surveys that literature and suggests that researchers in that field have not yet developed a set of best practices for the use of CVM with the types of choices that are most common in health economics contexts.

In the transportation and marketing literatures, the desired demand information often spanned a number of possible alternative products or services. Stated preference studies were often conducted not just to determine respondents' willingness to pay for a single well-defined good but to understand how willingness to pay might be affected by variations in the mix of attributes making up a prospective good. It was often necessary to anticipate demands for differentiated products, where each product could be characterized as a bundle of attributes and the levels of these attributes differed across alternatives.

In contrast, more of the impetus for CVM non-market valuation research in environmental economics stemmed from a number of significant lawsuits. In the legal context, there is a premium on simplicity in economic modelling so as not to confuse the jury or the judge. It is often best to produce one value for one clearly defined commodity. (Providing a judge or jury with a function that describes demand, where WTP depends upon a wide array of attributes, conditions or respondent attributes, can actually be a liability when attorneys are trying to make a simple, clear and persuasive case. In a legal context, it is most incisive to value one thing, and to value it as precisely as possible.) Eventually, however, environmental economists began to acknowledge the value of understanding the heterogeneity in demands for environmental goods, since this knowledge can be very helpful to policymakers who wish to consider how different versions of a policy might affect different constituencies.

There are many commonalities between the tasks faced by environmental economists and those faced by transportation economists and market researchers, but there is one key difference. In transportation economics and market research, it is often the case that the public transit system in question will actually be built, or the new product will actually be developed and put on the market. There is an opportunity to go back and see whether the level of demand predicted by the stated preference study actually materializes when there is a real market. In the environmental

economics literature, there are typically fewer opportunities to 'validate' the stated preference demand information with revealed preferences for the same product.

One common expectation for a good CVM study is now a demonstration that the demand function that has been estimated should 'walk and talk' like a demand function. For example, is willingness to pay to preserve big-game hunting opportunities lower, on average, for elderly women than for middle-aged males? Is willingness to pay to preserve air quality higher for people with lung disease or asthma, or for people who have family members with these illnesses? These tests are commonly called 'construct validity' assessments. Contingent valuation studies that pass a battery of plausibility tests such as these can generally be viewed as more reliable.

Another common test of contingent valuation estimates that these stated preference demand functions are typically expected to satisfy is something called a 'scope test'. This means that, on average, respondents' willingness to pay for an alternative that involves more of the 'good' in question should be greater than that for an alternative that involves less of the 'good'. It is of course possible that marginal utility may be positive (as the scope test implies) at low levels of the good, but also that it may go to zero if the quantity of the good is high enough for satiation to set in, and there is no theoretical basis for expecting willingness to pay to be proportional to the amount of the good in question.

CVM data can also sometimes be pooled with actual choice data. This can allow portions of the underlying indirect utility function to be anchored upon real choices, even though the variability in attributes in the real alternatives may not span the full domain relevant to pending policy decisions. The stated choice questions can be used to extend the domain of the estimated demand function.

While economists will remain uncomfortable about reliance upon stated preference information, many now acknowledge that there are circumstances where stated preference data are all that can be collected. In fact, the need for economists to rely upon survey data (what people say as opposed to what they actually do in markets) is now being acknowledged in the other contexts in economics. For example, expectations about future income or life expectancy figure prominently in a number of economic theories. These expectations typically cannot be measured directly, but can sometimes be elicited using surveys and put to good use empirically (see Manski 2004).

It is worth noting that not just stated preference data but also revealed preference data can be highly variable in its quality. Much revealed preference demand data is also drawn from consumer surveys. It is not always clear that the individual respondent sees the need for accuracy and completeness to be as critical as researchers using the data might hope. In consumer expenditure surveys, for example, interviewers prompt subjects for different types of expenditures, but the enthusiasm and engagement of the survey subject often determines the accuracy and completeness of the data. Rather than viewing revealed preference data as of unambiguously high quality and stated preference data (such as that produced by CVM) as of unambiguously low quality, it may be prudent simply to acknowledge that both types of data can have their problems.

A partial list of current frontiers in CVM-related research is possible. These frontiers include continuing assessment of (*a*) alternative elicitation formats (there are many candidates beyond the simple NOAA-recommended binary choice format), (*b*) the choice contexts presented to subjects, (*c*) the effects of allowing subjects to express uncertainty about their choices, (*d*) the effects of practice and fatigue when several CVM questions are presented to each respondent, (*e*) integrating stated choices with additional types of real market information, (*f*) how the degree of complexity of the CVM choice scenarios interacts with the cognitive capacity of the subject and/or the subject's inclination to pay attention, and (g) the neuroeconomics of real as opposed to stated choices.

Two of the classic books on CVM are Cummings et al. (1986) and Mitchell and Carson (1989). Following the Exxon Valdez case, a provocative debate was featured in the *Journal of Economic Perspectives* (Diamond and Hausman 1994; Hanemann 1994; Portney 1994). McFadden (1994) raised some specific concerns about the reliability of CVM data in the context of an empirical application to the existence value of wilderness areas in the western United States. In the intervening years, however, research concerning CVM has continued apace. Helpful expositions and discussions of recent innovations have made their way into textbook form, with one particularly useful summary being provided in Chapter 6 of Freeman (2003). A brief, accessible and very helpful introduction to CVM for non-specialists is contained in Carson (2000). Louviere et al. (2000) offer a comprehensive discussion of stated choice methods broadly defined, including experimental design, modelling, estimation and combining revealed and stated preference data, with illustrations in marketing, transportation and environmental economics. An inventory of the wide range of practical issues to consider in actually implementing a CVM study is provided by Boyle (2003), while Holmes and Adamowicz (2003) update the state of the art for attribute-based (conjoint choice) methods.

There is still considerable variation in individual researchers' levels of comfort with CVM and stated preference data more generally. We might reconsider the question posed by Diamond and Hausman (1994): 'Contingent valuation – is some number better than no number?' There are now many economists who would agree that a value based on stated preferences – from a study that is carefully conceived and executed, based on a sufficiently large sample that is representative of its intended population, that has been put through a battery of consistency and validity assessments, and that produces an implied demand function that behaves the way we would expect a 'real' demand function to behave – is almost certainly better than no number. This is especially true when 'no number' creates the risk that a value of zero would otherwise be imputed, by default, for use in policy decisions.

See Also

► Environmental Economics

Bibliography

Arrow, K., R. Solow, P.R. Portney, E.E. Leamer, R. Radner, and H. Schuman. 1993. Report of the NOAA panel on contingent valuation. *Federal Register* 58: 4601–4614.

Boyle, K.J. 2003. Contingent valuation in practice. In *A primer on nonmarket valuation*, ed. P.A. Champ, K.J. Boyle, and T.C. Brown. Boston: Kluwer Academic Publishers.

Carson, R.T. 2000. Contingent valuation: A user's guide. *Environmental Science & Technology* 34: 1413–1418.

Carson, R.T., N.E. Flores, K.M. Martin, and J.L. Wright. 1996. Contingent valuation and revealed preference methodologies: Comparing the estimates for quasi-public goods. *Land Economics* 72: 80–99.

Cummings, R.G., D.S. Brookshire, and W.D. Schulze, eds. 1986. *Valuing environmental goods: An assessment of the contingent valuation method*. Totowa: Rowman and Allanheld.

Diamond, P.A., and J.A. Hausman. 1994. Contingent valuation – Is some number better than no number? *Journal of Economic Perspectives* 8(4): 45–64.

Freeman, A.M.I. 2003. *The Measurement of Environmental and Resource Values: Theory and Methods*. Washington, DC: Resources for the Future.

Hanemann, W.M. 1994. Valuing the environment through contingent valuation. *Journal of Economic Perspectives* 8(4): 19–43.

Holmes, T.P., and W.L. Adamowicz. 2003. Attribute-based methods. In *A primer on nonmarket valuation*, ed. P.A. Champ, K.J. Boyle, and T.C. Brown. Boston: Kluwer Academic Publishers.

Louviere, J.J., D.A. Hensher, and J.D. Swait. 2000. *Stated choice methods*. New York: Cambridge University Press.

Manski, C.F. 2004. Measuring expectations. *Econometrica* 72: 1329–1376.

McFadden, D. 1994. Contingent valuation and social choice. *American Journal of Agricultural Economics* 76: 689–708.

Mitchell, R.C., and R.T. Carson. 1989. *Using surveys to value public goods: The contingent valuation method*. Washington, DC: Resources for the Future.

Portney, P.R. 1994. The contingent valuation debate – Why economists should care. *Journal of Economic Perspectives* 8(4): 3–17.

Smith, R.D. 2003. Construction of the contingent valuation market in health care: A critical assessment. *Health Economics* 12: 609–628.

Continuity in Economic History

Donald N. McCloskey

Continuity and discontinuity are devices of story-telling, telling the story of monetary policy over the past few months or the story of modern economic growth. They raise certain questions in philosophy and lesser matters, such as precedence and politics.

It is well to have a case in mind. The most important is that of the British industrial revolution.

If it was a 'revolution', as it surely was, it happened sometime. There was a discontinuity, a before and after. When? Various dates have been proposed, down to the day and year: 9 March 1776, when the *Wealth of Nations* provided an ideology for the age; the five months in 1769 when Watt took out a patent on the high pressure steam engine and Arkwright on the cotton-spinning water frame; or 1 January 1760, when the furnaces at Carron Ironworks, Stirlingshire, were lit.

Such dating has of course an amateur air. A definite date looks handsome on a plaque or scroll but the precision does not fit well with sophisticated story-telling. The discontinuity is implausibly sharp, drawing attention to minor details. The Great Depression did not start on 24 October 1929; the deregulation of American banking was not completed with the fall of Regulation Q. Nicholas Crafts (1977) has pointed out that the detailed timing of the industrial revolution should not anyway be the thing to be studied, because small beginnings do not come labelled with their probabilities of developing into great revolutions. He is identifying a pitfall in story-telling. Joel Mokyr identifies another (1985, p. 44): rummaging among the possible acorns from which the great oak of the industrial revolution grew 'is a bit like studying the history of Jewish dissenters between 50 BC and 50 AD. What we are looking at is the inception of something which was at first insignificant and even bizarre', though 'destined to change the life of every man and woman in the West'.

What is destined or not destined to change our lives will look rather different to each of us. Each historian therefore has his or her own dating of the industrial revolution. Each sees another discontinuity. E.M. Carus-Wilson (1941, p. 41) spoke of 'an industrial revolution of the 13th century': she found that the fulling mill was 'due to scientific discoveries and changes in technique' and 'was destined to alter the face of medieval England'. A.C. Bridbury (1975, p. xix–xx) found in the late middle ages 'a country travelling slowly along the road … that [it] travelled so very much more quickly in Adam Smith's day'. In the eyes of Marxist writers the 16th century was the century of discontinuity, when capitalism set off into the world to seek its fortune. John U. Nef, no Marxist, believed he saw an industrial revolution in the 16th century, centred on coal (1932), though admittedly slowed in the 17th century. A student of the 17th century itself, such as D.C. Coleman (1977), finds glimmerings of economic growth even in that disordered age. The most widely accepted period for the industrial revolution is the late 18th century, especially the 1760s and 1770s (Mantoux 1928; Landes 1969), but recent students of the matter (Harley 1982; Crafts 1984) have found much to admire in the accomplishments of the early 18th century. W.W. Rostow (1960) placed the 'takeoff into self-sustained growth' in the last two decades of the 18th century, but others have observed that even by 1850 the majority of British people remained in traditional sectors of the economy. And later still there was a second industrial revolution (of chemicals, electricity, and internal combustion) and a third (of electronics and biology).

Wider perspectives are possible, encouraging the observer to see continuity instead. Looking at the matter from 1907, the American historian Henry Adams could see a 'movement from unity into multiplicity, between 1200 and 1900, …

unbroken in sequence, and rapid in acceleration' (p. 498). The principal modern student of the industrial revolution, R.M. Hartwell, appealed for continuity against the jostling throng of dates (1965, p. 78): 'Do we need an *explanation* of the industrial revolution? Could it not be the culmination of a most unspectacular process, the consequence of a long period of economic growth?'

Such questions of continuity and discontinuity are asked widely in economics, though sometimes half consciously. They should not be left to historians. Economics is mainly contemporary history, and faces the problem of deciding when a piece of history has been continuous or not. For instance the crucial discontinuity in the growth of big government, as Robert Higgs (1987) points out, might be placed when the institutions of centralized intervention were conceived (1900–1918) or made (1930–45) or expanded (1960–70). Even recent history faces this narrative problem. When, if ever, did purchasing power parity break down in the 1970s? When did policy on antitrust alter to favour mergers? When did monetary policy last become expansionary? Where is the break?

The difficulty in answering the question has often been misconstrued as philosophical. The philosophical difficulty was first articulated in the 5th century BC by Parmenides and his student Zeno: that if everything is perfectly continuous, change is impossible (Korner 1967). Everything is so to speak packed too tightly to move. The economist will recognize the point as analogous to an extreme form of economic equilibrium, or to the physicist's maximum entropy. If human nature doesn't 'really' change, then history will be a string of weary announcements that the more things change the more they stay the same. If the economy is 'really' in equilibrium all the time, then nothing remains to be done.

Alexander Gerschenkron, the economic historian who has contributed most to the understanding of continuity and discontinuity in economics, noted that such a metaphysics would close the book of history (1962, p. 12). A history of economics that began with the Parmenidean continuum would never speak.

For purposes of social science Gerschenkron rejects the transition from the connectedness of all change to an absence of change. True, if you squint and fit a curve then no economic change looks discontinuous in the mathematical sense; but it is wrong then to deduce that 'really' there is no change at all, or that the industrial revolution is a mirage. 'Continuity' in the strict mathematical sense must be kept distinct from 'continuity' in the story-telling sense.

Economists have often been muddled about this philosophical distinction, drawing surprising ideological implications from it. Alfred Marshall enshrined on the title page of his *Principles* the motto 'natura non facit saltum' (nature does not make a jump; Leibnitz had invented it as 'la nature ne fait jamais des sauts'). Marshall himself perhaps believed that the ability to represent behaviour with differentiable functions implies that marginalism is a good description of human behaviour. It is less sure that he believed that the lack of jumps in nature (this on the eve of quantum physics) implies people should not jump either, and should change society only gradually. Anyway, both implications are non sequiturs. Though both have been attributed to neoclassical economics, neither is necessary for it. Much bitter controversy has assumed that neoclassical economics depends on smooth curves and in consequence must advocate smooth social policies. The peculiar alliance between discrete mathematics and Marxian economics has this origin, as does the enthusiasm of some conservative writers for continuities in economic history. Gerschenkron cursed both their houses; the social scientist should study change and continuity 'unbothered by the lovers and haters of revolutions who must find themselves playgrounds and battlegrounds outside the area of serious scholarship' (p. 39).

In one sense of 'continuity' it is trivial that economic history is continuous. History has causes (the fourth of five historically relevant definitions that Gerschenkron distinguishes). Continuity, then, can be viewed as being merely an impressively long causal chain. The exploitation of Scottish iron deposits in the 18th century was caused by bold investments, but these depended on a reliable law of property and commerce, which depended on certain legal developments in the 16th century, and on the growth of

political stability in the early 18th century, which in turn depended on all manner of earlier events. Establishing continuities, as Gerschenkron remarks, is the historian's purpose – or, one might add, the economist's, who is doing historian's work when he is not doing philosopher's. The purpose might be to find a cause of, say, the Great Depression. It would be to find a chain of events the absence of which would have made a difference: the international irresponsibility of the United States, for instance, as Kindleberger argued; or the domestic irresponsibility of the Federal Reserve, as Friedman and Schwartz argued. Finding such chains has its own philosophical difficulties (see the article in this Dictionary on "► Counterfactuals)".

The main problems of continuity and discontinuity, however, are not solvable in seminars on philosophy. They are practical problems in the uses of measurement, and must be solved in the economic or historical workshop. When shall we say that the industrial revolution happened? Gerschenkron gives an answer confined to industry, for in common with most economic historians he regards agriculture and services as laggards in economic growth.

> In a number of major countries of Europe . . . after a lengthy period of fairly low rates of growth came a moment of more or less sudden increase in the rates, which then remained at the accelerated level for a considerable period. That was the period of the great spurt in the respective countries' industrial development . . .. The rates and the margin between them in the 'pre-kink' and the 'post-kink' periods appear to vary depending on the degree of relative backwardness of the country at the time of the acceleration. (pp. 33–4)

The level at which such discontinuity is to be observed is at choice. As Gerschenkron remarks,

> If the seat of the great spurt lies in the area of manufacturing, it would be inept to try to locate the discontinuity by scrutinizing data on large aggregate magnitudes such as national income . . .. By the time industry has become bulky enough to affect the larger aggregate, the exciting period of the great spurt may well be over. (pp. 34–5)

In a footnote to these sentences he remarks that 'Walt Rostow's failure to appreciate this point has detracted greatly from his concept of the take-off, which in principle is closely related to the concept of the great spurt as developed by this writer'.

The point is a good one, and applies to all questions of continuity in aggregate economics. Small (and exciting) beginnings will be hidden by the mass until well after they have become routine. Joel Mokyr has put it as a matter of arithmetic: if the traditional sector of an economy is growing at a slow one per cent per annum, and starts with 90 per cent of output, the modern sector growing at four per cent per annum will take three-quarters of a century to account for as much as half of output (1985, p. 5). We may call it the Weighting Theorem (or the Waiting Theorem, for the wait is long when the weight is small to begin with). There are parallel points to be made elsewhere in economics and in social science generally. In growth theory, for instance, as was noticed shortly after its birth, a century of theoretical time is needed in most models for a shift to yield growth as much as 90 per cent of its steady state. More generally, economists have long recognized the tension between microeconomic explanations and the macroeconomic things to be explained. And sociologists have been quarrelling along similar lines for a century, using even the same jargon of micro and macro.

In other words, the search for discontinuity in an aggregate time series raises the question of the level at which we should do our social thinking, the aggregation problem. Yet Gerschenkron himself did not answer the question well, and was hoist by his own petard. Calculating Italian industrial output he placed his 'big spurt' in 1896–1908, and wished to explain it with big banks founded in the 1890s. Stefano Fenoaltea, once his student, applied the Weighting Theorem to the case (Fenoaltea 1987). Surely, Fenoaltea reasoned, the components of the industrial index – the steel output and the chemical output – are the 'real' units of economic analysis (note the similarity of this rhetoric to that advocating a micro foundation for macroeconomics). If the components started accelerating *before* the new banks appeared, becoming bulky only later, then the new banks could not have been the initiating force. Alas, the components did just this. They spoil Gerschenkron's bank-led story: the

components accelerated not in the 1890s but in the 1880s, not after but before the banks. To paraphrase Gerschenkron on Rostow, by the time the progressive components of industry had become bulky enough to affect the larger aggregate, the exciting period was well over.

Yet the moral is still Gerschenkron's: that continuity and discontinuity are tools 'forged by the historian rather than something inherently and invariantly contained in the historical matter . . .. [A]t all times it is the ordering hand of the historian that creates continuities or discontinuities' (p. 38). Gerschenkron nodded, but in nodding made the point. The multiple datings of the industrial revolution make it, too. So does any choice of smoothness or suddenness in economic storytelling.

The point is that history, like economics, is a story we tell. Continuity and discontinuity are narrative devices, to be chosen for their storytelling virtues. Niels Bohr said once that 'It is wrong to think that the task of physics is to find out how nature is. Physics concerns what we can say about nature'. It is *our* say. We can choose to emphasize the continuous: 'Abraham begat Isaac; . . . begat . . . begat . . . and Jacob begat Joseph the husband of Mary, of whom was born Jesus'. Or the discontinuous: 'There was in the days of Herod, the king of Judea, a certain priest named Zacharias.' It is the same story, but its continuity or discontinuity is our creation, not God's. That it is out of God's hands does not make it arbitrary. Scholars speak of the industrial revolution as early or late, gradual or sudden. Other scholars believe or disbelieve their stories on the usual grounds.

See Also

► Economic History

Bibliography

Adams, H. 1906. *The education of Henry Adams*. New York: Modern Library. 1931.

Bridbury, A.C. 1975. *Economic growth: England in the later middle ages*. Brighton: Harvester.

Carus-Wilson, E.M. 1941. An industrial revolution of the thirteenth century. *Economic History Review* 11(1): 39–60. Reprinted in *Essays in economic history*, vol. I, ed. E.M. Carus-Wilson. London: Edward Arnold, 1954.

Coleman, D.C. 1977. *The economy of England 1450–1750*. Oxford: Oxford University Press.

Crafts, N.F.R. 1977. Industrial revolution in England and France: Some thoughts on the question 'Why was England first?'. *Economic History Review, 2nd series* 30(3): 429–441.

Crafts, N.F.R. 1984. *Economic growth during the British industrial revolution*. Oxford: Oxford University Press.

Fenoaltea, S. 1987. *Italian industrial production, 1861–1913: A statistical reconstruction*. Cambridge: Cambridge University Press.

Gerschenkron, A. 1962. On the concept of continuity in history. *Proceedings of the American Philosophical Society*, June. Reprinted in A. Gerschenkron, *Continuity in history and other essays.* Cambridge, MA: Harvard University Press, 1968.

Harley, C.K. 1982. British industrialization before 1841: Evidence of slower growth during the industrial revolution. *Journal of Economic History* 42(2): 267–290.

Hartwell, R.M. 1965. The causes of the industrial revolution: an essay in methodology. *Economic History Review, 2nd series* 18: 164–182. Reprinted in *The causes of the industrial revolution in England*, ed. R.M. Hartwell. London: Methuen, 1967.

Higgs, R. 1987. *Crisis and Leviathan: Critical episodes in the growth of American government*. New York: Oxford University Press.

Korner, S. 1967. Continuity. In *The encyclopedia of philosophy.* New York: Macmillan and Free Press.

Landes, D.S. 1969. *The unbound prometheus*. Cambridge: Cambridge University Press.

Mantoux, P. 1928. *The industrial revolution in the eighteenth century.* New York: Harper, 1961.

Mokyr, J. (ed.). 1985. *The economics of the industrial revolution*. Totowa: Rowman and Allanheld.

Nef, J.U. 1932. *The rise of the British coal industry,* 2 vols. London: Routledge.

Rostow, W.W. 1960. *The stages of economic growth*. Cambridge: Cambridge University Press.

Continuous and Discrete Time Models

Christopher A. Sims

Abstract

Most modelling of economic time series works with discrete time, yet time is in fact continuous. While in many instances simple intuitive

connections exist between results with discrete time data and the underlying continuous time dynamics, it is possible for discretization to create bias or have unintuitive effects. Some economics literature investigates such distortions. It is also possible to estimate explicitly continuous-time models, using discrete data. This approach raises its own difficulties, but has become more usable as computing power and the techniques to exploit it have improved.

Keywords

Approximation theory; Continuous and discrete time models; Distributed lags; Dynamic stochastic general equilibrium models; Granger causal priority; Markov chain Monte Carlo methods; Martingales; No-arbitrage models; Stochastic differential equations; Vector autoregressions; Wiener process

JEL Classifications

C3

Discrete time models are generally only an approximation, and the error induced by this approximation can under some conditions be important.

Most economists recognize that the use of discrete time is only as an approximation, but assume (usually implicitly) that the error of approximation involved is trivially small relative to the other sorts of simplification and approximation inherent in economic theorizing. We consider below first the conditions under which this convenient assumption may be seriously misleading. We discuss briefly how to proceed when the assumption fails and the state of continuous time economic theory.

Approximation Theory

Some economic behaviour does involve discrete delays, and most calculated adjustments in individual patterns of behaviour seem to occur following isolated periods of reflection, rather than continually. These notions are sometimes invoked to justify economic theories built on a discrete time scale. But to say that there are elements of discrete delay or time discontinuity in behaviour does not imply that discrete time models are appropriate. A model built in continuous time can include discrete delays and discontinuities. Only if all delays were discrete multiples of a single underlying time unit, and synchronized across agents in the economy, would modelling with a discrete time unit be appropriate.

Nonetheless, sometimes discrete models can avoid extraneous mathematical complexity at little cost in approximation error. It is easy enough to argue that time is in fact continuous and to show that there are in principle cases where use of discrete time models can lead to error. But it is also true in practice that more often than not discrete time models, translated intuitively and informally to give implications for the real continuous time world, are not seriously misleading. The analytical task, still not fully executed in the literature, is to understand why discrete modelling usually is adequate and thereby to understand the special circumstances under which it can be misleading.

The basis for the usual presumption is that, when the time unit is small relative to the rate at which variables in a model vary, discrete time models can ordinarily provide good approximations to continuous time models. Consider the case, examined in detail in Geweke (1978), of a dynamic multivariate distributed leg regression model, in discrete time.

$$Y\ (t) = A^{*}X(t) + U(t), \tag{1}$$

where * stands for convolution, so that

$$A^{*}X(t) = \sum_{s=-\infty}^{\infty} A(s)X(t-s). \tag{2}$$

We specify that the disturbances are uncorrelated with the independent variable vector X, that is, $\mathrm{cov}[X(t), U(s)] = 0$, all t, s. The natural assumption is that, if approximation error from use of discrete time is to be small, $A(s)$ must be

smooth as a function of s, and that in this case (1) is a good approximation to a model of the form

$$y(t) = a^*x(t) + u(t) \tag{3}$$

where

$$a^*x(t) = \int_{-\infty}^{\infty} a(s)x(t-s)ds \tag{4}$$

and y, a and x are functions of a continuous time parameter and satisfy $y(t) = Y(t)$, $x(t) = X(t)$ and $a(t) = A(t)$ at integer t. In this continuous time model we specify, paralleling the stochastic identifying assumption in discrete time, $\text{cov}[x(t), u(s)] = 0$, all t, s. If the discrete model (2) corresponds in this way to a continuous time model, the distributed lag coefficient matrices A(s) are uniquely determined by a and the serial correlation properties of x.

We should note here that, though this framework seems to apply only to the case where X is a simple discrete sampling of x, not to the time-averaged case where $X(t)$ is the integral of $x(s)$ from $t-1$ to t, in fact both cases are covered. We can simply redefine the x process to be the continuously unit-averaged version of the original x process. This redefinition does have some effect on the nature of limiting results as the time unit goes to zero (since the unit-averaging transformation is different at each time unit) but turns out to be qualitatively of minor importance.

Roughly speaking, sampling a unit-averaged process is like sampling a process whose paths have derivatives of one higher order than the unaveraged process.

Geweke shows that under rather general conditions

$$\sum_{s=-\infty}^{\infty} \|A(s) - \tau a(s\tau)\|^2 \to 0 \tag{5}$$

as the time unit τ goes to zero, where $\| \ \|$ is the usual root-sum-of-squared-elements norm. In this result, the continuous time process x and lag distribution a are held fixed while the time interval corresponding to the unit in the discrete time model shrinks.

This is the precise sense in which the intuition that discrete approximation does not matter much is correct. But there are important limitations on the result. Most obviously, the result depends on a in (3) being an ordinary function. In continuous time, well-behaved distributed lag relations like (3) are not the only possible dynamic relation between two series. For example, if one replaces (3) by

$$y(t) = \alpha(\mathrm{d}/\mathrm{dt})x(t) + u(t), \tag{6}$$

then the limit of A in (2) is different for different continuous x processes. In a univariate model with second-order Markov x (for example, one with $\text{cov}\ [x(t), x(t-s)] = (1+\theta|s|)\mathrm{e}^{-\theta|s|}\ \text{var}[x(t)]$, the limiting discrete time model, as τ goes to zero, is

$$\begin{aligned} y(t) = \ & a\{-0.02X(t+4) + 0.06X(t+3) - 0.22X(t+2) \\ & + 0.80X(t+1) - 0.80X(t-1) + 0.22X(t-2) \\ & - 0.06X(t-3) + 0.02X(t-4)\} + U(t) \end{aligned} \tag{7}$$

(see Sims 1971).

This result is not as strange as it may look. The coefficients on X sum to zero and are antisymmetric about zero. Nonetheless, (7) is far from the naive approximation which simply replaces the derivative operator with the first difference operator. In fact, if the estimation equation were constrained to involve only positive lags of X, the limiting form would be

$$\begin{aligned} Y(t) = \alpha\{ & 1.27X(t) - 1.161X(t-1) \\ & + 0.43X(t-2) - 0.12X(t-3) \\ & + 0.03X(t-4) - 0.01X(t-5)\} + U(t). \end{aligned} \tag{8}$$

The naive approximation of (3) by $Y(t) = \alpha[X(t) - X(t-1)] + U(t)$ is valid only in the sense that, if this form is imposed on the discrete model a priori, the least squares estimate of α will converge to its true value. If the resulting estimated model is tested for fit against (8) or (7), it will be rejected.

Although the underlying model involves only the contemporaneous derivative of x, (8) and (7) both involve fairly long lags in X. If x paths have

higher than firstorder derivatives (for example, if they are generated by a third-order stochastic differential equation) the lag distributions in (8) and (7) are replaced by still higherorder limiting forms. Thus, different continuous time processes for x which all imply differentiable time paths produce different limiting discrete A. Here the fact that the time unit becomes small relative to the rate of variation in x does not justify the assumption that approximation of continuous by discrete models is innocuous. In particular, the notion that discrete differencing can approximate derivatives is potentially misleading.

It should not be surprising that the discrete time models may not do well in approximating a continuous time model in which derivatives appear. Nonetheless, empirical and theoretical work which ignores this point is surprisingly common.

If a is an ordinary function, there is still chance for error despite Geweke's result. His result implies only that the mean square deviation of a from A is small. This does not require that individual $A(t/\tau)$'s converge to the corresponding $a(t)$ values. For example, in a model where x is univariate and $a(t) = 0, t < 0, a(0) = 1, a(s)$ continuous on $[0,\infty]$, the limiting value for $A(0)$ is 0.5, not 1.0. Thus, if $a(t) = e^{-\theta tt}$ on $[0, \infty)$, making a monotone decreasing over that range, $A(t)$ will not be monotone decreasing. It will instead rise between $t = 0$ and $t = 1$. This is not unreasonable on reflection: the discrete lag distribution gives a value at $t = 0$ which averages the continuous time distribution's behaviour on either side of $t = 0$. It should therefore not be surprising that monotonicity of a does not necessarily imply monotonicity of A, but the point is ignored in some economic research.

Another example of possible confusion arises from the fact that, if the x process has differentiable paths, $a(t) = 0$ for $t < 0$ does not imply $A(t) = 0$ for $t < 0$. The mean-square approximation result implies that when the time unit is small the sum of squares of coefficients on $X(t - s)$ for negative s must be small relative to the sum of squares on $X(t - s)$ for positive s, but the first few lead coefficients will generally be non-zero and will not go to zero as the time interval goes to zero. This would lead to mistaken conclusions about Granger causal priority in large samples, if significance tests were applied naively.

Geweke's exploration of multivariate models shows that the possibilities for confusing results are more numerous and subtle in that case. In particular, there are ways by which poor approximation of $\alpha_j(s)$ by $A_j(s/\tau)$ in some s interval (for example, around $s = 0$) can lead to contamination of the estimates of other elements of the A matrix, even though they correspond to x_j's and a_j's that in a univariate model would not raise difficulties.

In estimation of a dynamic prediction model for a single vector y, such as a vector autoregression (VAR) or dynamic stochastic general equilibrium model (DSGE), the question for approximation theory becomes whether the continuous time dynamics for y, summarized in a Wold moving average representation

$$y(t) = a^* u(t) \tag{9}$$

has an intuitively transparented connection to the corresponding discrete time Wold representation

$$Y(t) = A^* U(t). \tag{10}$$

In discrete time the $U(t)$ of the Wold representation is the one-step-ahead prediction error, and in continuous time $u(t)$ also represents new information about y arriving at t. There are two related sub-questions. Is the A function the same shape as the a function; and is the U vector related in a natural way to the u vector? The u vector is a continuous time white noise, so that U cannot possibly be a simple discrete sampling of u.

If y is stationary and has an autoregressive representation, then $U(t) = A^{-1} * a * u_t$, with the expression interpreted as convolution in continuous time, but with A^{-1} putting discrete weight on integers. The operator connecting U and u is then $A^{-1} * a$. There are cases where the connection between continuous and discrete time representations is intuitive. For example, if $a(s) = \exp(-Bs)$ (with the exponentiation interpreted as a matrix exponential in a multivariate case), then

$$U(t) = \int_0^1 e^{-Bs} u(t-s)ds \tag{11}$$

and $A(s) = a(s)$ at integers. This is a more intuitive and precise matching than in any case we examined above for projection of one variable on another. If $a(0)$ is full rank and right-continuous at zero and if $a(s)$ is differentiable at all $s > 0$, then a similar intuitively simple matching of A to a arises when the time unit is small enough.

However, non-singularity of $a(0)$ rules out differentiability of time paths for y. When time paths for y, or some elements of it, are differentiable, no simple intuitive matching between A and a arises as the time unit shrinks.

There is one clear pattern in the difference in shape between A and a that stands in contrast to the case of distributed lag projection considered above. If both the continuous time and the discrete time moving average representations are fundamental, then by definition the one-step-ahead prediction error in $y(t)$ based on $y(t-s)$, $s \geq 1$ is

$$\int_0^1 a(s)u(t-s)ds, \tag{12}$$

while the one-step-ahead prediction error in $Y(t)$ based on $Y(t-s)$, $s = 1, 2,...$ is $A_0 Y(t)$. Now the information set we use in forecasting based on the past of Y at integer values alone is smaller than the information set based on all past values of y, so the one-step-ahead error based on the discrete data alone must be larger. If we normalize in the natural way to give U an identity covariance matrix and to make var $(g*u(t)) = \int g(s)g'(s)$ (so u is a unit white noise vector), then it must emerge that

$$A_0 A_0 \geq \int_0^1 a(s)a(s)' ds, \tag{13}$$

where the inequality is interpreted as meaning that the left-hand-side matrix minus the right-hand-side matrix is positive semi-definite. In other words, the initial coefficient in the discrete MAR will always be as big or bigger than the average over (0,1) of the coefficients in the continuous MAR. This tendency of the discrete MAR to seem to have a bigger instant response to innovations is proportionately larger the smoother a is near zero.

More detailed discussion of these points, together with numerous examples, appears in Marcet (1991).

Estimation and Continuous Time Modelling

How can one proceed if one has a model like, say, (6), to which a discrete time model is clearly not a good approximation? The only possibility is to introduce explicitly a model for how x behaves between discrete time intervals, estimating this jointly with (6) from the available data. Doing so converts (6) from a single-equation to a multiple-equation model. That is, the device of treating x as 'given' and non-stochastic cannot work because an important part of the error term in the discrete model arises from the error in approximating $a*x$ by $A*X$. Furthermore, because separating the approximation error component of U from the component due to u is essential, one would have to model serial correlation in u explicitly. The model could take the form

$$\begin{bmatrix} y(t) \\ x(t) \end{bmatrix} = \begin{bmatrix} c(s) & a^* b(s) \\ 0 & b(s) \end{bmatrix} * \begin{bmatrix} w(t) \\ v(t) \end{bmatrix}, \tag{14}$$

where w and v are white noise processes fundamental (in the terminology of Rosanov 1967), for y and x. To give b and c a convenient parametric form, one might suppose them rational, so that (14) can be written as a differential equation, that is,

$$P(D)\, y(t) = P(D)a^* x(t) + w(t) \tag{15}$$

$$Q(D)x(t) = v(t), \tag{16}$$

where P and Q are finite-order polynomials in the derivative operator, $Q^1(D)v = b*v$, and $P^{-1}(D)w = c*w$.

A discrete time model derived explicitly from a continuous time model is likely to be nonlinear at

least in parameters and therefore to be more difficult to handle than a more naive discrete model. However with modern computing power, such models are usable. Bergstrom (1983) provides a discussion of estimating continuous time constant coefficient linear stochastic differential equation systems from discrete data, the papers in the book (1976) he edited provide related discussions, and Hansen and Sargent (1991), in some of their own chapters of that book, discuss estimation of continuous time rational expectations models from discrete data.

Estimating stochastic differential equation models from discrete data has recently become easier with the development of Bayesian Markov chain Monte Carlo (MCMC) methods. Though implementation details vary across models, the basic idea is to approximate the diffusion equation

$$dy_t = a(y_t)dt + b(y_t)dW_t, \tag{17}$$

where W_t is a Wiener process, by

$$y_t = e^{-a(y_{t-\delta})}y_{t-\delta} + b(y_{t-\delta})\varepsilon_t. \tag{18}$$

Such an approximation can be quite inaccurate unless δ is very small. But one can in fact choose δ very small, much smaller than the time interval at which data are observed. The values of y_t at times between observations are of course unknown, but if they are simply treated as unknown 'parameters it may be straightforward to sample from the joint posterior distribution of the y's at non-observation times and the unknown parameters of the model. The Gibbs sampling version of MCMC samples alternately from conditional posterior distributions of blocks of parameters. Here, sampling from the distribution of y at non-observation dates conditioning on the values of model parameters is likely to be easy. If the model has a tractable form, it will also be easy to sample from the posterior distribution of the parameters conditional on all the y values, both observed and unobserved. Application of these general ideas to a variety of financial models is discussed in Johannes and Polson (2006).

Another approach that has become feasible with increased computing power is to develop numerical approximations to the distribution of $y_{t+\delta}$ conditional on data through time t. Ait-Sahalia (2007) surveys methods based on this approach.

Modelling in continuous time does not avoid the complexities of connecting discrete time data to continuous time reality – it only allows us to confront them directly. One reason this is so seldom done despite its technical feasibility is that it forces us to confront the weakness of economic theory in continuous time. A model like (15)–(16) makes an assertion about how many times y and x are differentiable, and a mistake in that assertion can result in error as bad as the mistake of ignoring the time aggregation problem. Economic theory does not have much to say about the degree of differentiability of most aggregate macroeconomic time series. When the theory underlying the model has no believable restrictions to place on fine-grained dynamics, it may be better to begin the modelling effort in discrete time. As is often true when models are in some respect underidentified, it is likely to be easier to begin from a normalized reduced form (in the case the discrete time model) in exploring the range of possible interpretations generated by different potential identifying assumptions.

Recent developments in financial economics have produced one area where there are continuous time economic theories with a solid foundation. Stochastic differential equations (SDEs) provide a convenient and practically useful framework for modelling asset prices. These SDE models imply non-differentiable time paths for prices, and it is known (Harrison et al. 1984) that differentiable time paths for asset prices would imply arbitrage opportunities, if there were no transactions costs or bounds on the frequency of transactions.

However, there are in fact transactions costs and bounds on transactions frequencies, and no-arbitrage models for asset prices break down at very fine, minute-by-minute, time scales. Successful behavioural modelling of these fine time scales requires a good theory of micro-market structure, which is still work in progress.

It is worthwile noting that a process can have non-differentiable paths without producing white

noise residuals at any integer order of differentiation: for example, a model satisfying (3) with $a(s) = s^{0.5}e^{-s}$. Such a process has continuous paths with unbounded variation and is not a semimartingale. That is, it is not the sum of a martingale and a process with bounded variation, and therefore cannot be generated from an integer-order SDE. Similarly, if $a(s) = s^{0.5}e^{-s}$, the process has non-differentiable paths but is nonetheless not a semimartingale. The existence of such non-semimartingale processes and their possible applications to financial modelling is discussed in Sims and Maheswaran (1993).

See Also

► Time Series Analysis

Bibliography

Aït-Sahalia, Y. 2007. Estimating continuous-time models using discretely sampled data. In *Advances in economics and econometrics, theory and applications. Ninth World Congress*, ed. R. Blundell, T. Persson, and W.K. Newey, vol. 3. Cambridge: Cambridge University Press.

Bergstrom, A.R., ed. 1976. *Statistical inference in continuous time economic models*. Amsterdam: North-Holland.

Bergstrom, A.R. 1983. Gaussian estimation of structural parameters in higher order continuous time dynamic models. *Econometrica* 51: 117–152.

Geweke, J. 1978. Temporal aggregation in the multiple regression model. *Econometrica* 46: 643–662.

Hansen, L.P., and T.J. Sargent, eds. 1991. *Rational expectations econometrics*. Boulder and Oxford: Westview Press.

Harrison, J.M., R. Pitbladdo, and S.M. Schaefer. 1984. Continuous price processes in frictionless markets have infinite variation. *Journal of Business* 57: 353–365.

Johannes, M., and N. Polson. 2006. MCMC methods for continuous-time financial econometrics. In *Handbook of financial econometrics*, ed. Y. Aït-Sahalia and L.P. Hansen. Amsterdam: North-Holland.

Marcet, A. 1991. Temporal aggregation of economic time series. In *Rational expectations econometrics*, ed. L.P. Hansen and T.J. Sargent. Boulder and Oxford: Westview Press.

Rozanov, Yu.A. 1967. *Stationary random processes*, trans A. Feinstein. San Francisco/Cambridge/London/Amsterdam: Holden-Day.

Sims, C.A. 1971. Approximate specifications in distributed lag models. In *Proceedings of the 38th Session, Bulletin of the International Statistical Institute* 44, Book 1.

Sims, C.A., and S. Maheswaran. 1993. Empirical implications of arbitrage-free asset markets. In *Models, methods and applications of econometrics*, ed. P.C.B. Phillips. Oxford: Blackwell.

C

Continuous-Time Stochastic Models

Robert C. Merton

Models in which agents can revise their decisions continuously in time have proved fruitful in the analysis of economic problems involving intertemporal choice under uncertainty (cf. Malliaris and Brock 1982). These models frequently produce significantly sharper results than can be derived from their discrete-time counterparts. In the majority of such cases, the dynamics of the underlying system are described by diffusion processes, whose continuous sample paths can be represented by Ito integrals. However, in selected applications, this assumption can be relaxed to include both non-Markov path-dependent processes and Poisson-directed jump processes.

An early application of this mode of analysis was the lifetime consumption-portfolio selection problem (Merton 1969, 1971). Under the assumptions of continuous trading and asset returns generated by diffusion processes, the derived structure of optimal portfolio demands produce portfolio-separation or mutual fund theorems like those derived in the static Markowitz–Tobin mean-variance model, but without the objectionable assumption of either quadratic preferences or Gaussian-distributed asset prices. Indeed, in the special, but prototypical, case of lognormally-distributed asset prices, the intertemporal optimal rules are identical to those of the mean-variance model. The continuous-time analysis thus provides a reconciliation of this classic model, with models of general expected utility maximization

in an environment where asset ownership has limited liability.

Using these same assumptions of continuous trading and lognormality of security prices, Black and Scholes (1973) derived a formula for pricing options that provided the foundation for subsequent development, of a unified theory of corporation liability evaluation and general contingent-claim pricing. Cox et al. (1985a) use the continuoustrading methodology with diffusion processes to derive a general theory for the term structure of interest rates.

Building on the continuous-time model of individual choice, Merton (1973), Breeden (1979), and Cox et al. (1985b) develop intertemporal models of equilibrium asset prices. Huang (1985) provides a stronger foundation for these models by showing that if information in an economy with continuous-trading opportunities evolves according to diffusion processes, then equilibrium security prices will also evolve according to diffusion processes.

In the intertemporal version of the Arrow–Debreu general equilibrium model with complete markets, the markets need only to be open 'once' because agents will have no need for further trade. The continuous-trading model is in this respect at the opposite extreme. Economies in which the dynamics of the system are described by diffusion processes will have a continuum of possible states over any finite interval of time. Thus, in the strict sense, to have complete markets in the continuous-time diffusion model requires an uncountable number of pure Arrow–Debreu securities. The continuoustrading model with diffusions, nevertheless, appears to have many of the important properties of the Arrow–Debreu model, but without nearly so many securities.

As is well known, in the absence of complete Arrow–Debreu markets, a competitive equilibrium does not in general produce Pareto optimal allocations. However, Radner (1972) has shown that an Arrow–Debreu equilibrium allocation can be achieved without a full set of pure time-state contingent securities if agents can use the available securities to implement dynamic trading strategies which replicate the payoff structure of the missing pure securities. There is much analysis to suggest that continuous-trading opportunities together with diffusion representations sentations for the evolution of the economy provide a particularly fertile environment for fulfilling the Radner conditions.

Under reasonably general assumptions about agents' preferences and endowments Breeden (1979) among others has shown that the intertemporal equilibrium allocations generated in economies with continuous trading in a finite number of securities can be Pareto efficient. In the analysis of the individual portfolio selection problem underlying these equilibrium models, the derived portfolio-separation theorems show that the set of individually-optimal portfolios can be generated by combinations of relatively few composite securities or mutual funds.

The extensive literature on options and contingent-claims pricing provides further evidence that continuous-trading opportunities make possible a large reduction in the number of securities markets without loss of efficiency. Although typically partial equilibrium in nature, these analyses show that continuous-trading dynamic portfolio strategies using as few as two securities can replicate a wide range of state and time-dependent payoff structures.

In perhaps the most general analysis to date, Duffie and Huang (1985) study the role continuous trading plays in successfully implementing Arrow–Debreu equilibria with infinite dimensional commodity spaces, using only a finite number of securities. In particular, they derive necessary and sufficient conditions for continuous-trading portfolio strategies with a finite number of securities to effectively complete markets in a Radner economy. By working with martingale representation theorems, Duffie and Huang show that the class of dynamics for which these results obtain extends beyond vector diffusion processes to include some non-Markov path dependent processes. They also show that having heterogeneous probability assessments among agents provides no important difficulties with the results, provided all agents' subjective

probability measures are uniformly absolutely continuous. Although there remain further technical issues to be resolved, it is evident that the continuous-trading models provide a strong foundation for the belief that a good substitute for having many markets and securities is to have fewer markets which are open for trade more frequently.

A sketch of the derivation of the portfolio separation theorem along the lines of Merton (1971, 1973, 1982a) and Breeden (1979) is a follows:

At each time t, each consumer-investor acts so as to

$$\text{Max } E_t\left\{\int_t^T U[c(\tau), S(\tau), \tau]d\tau + B[W(T), S(T), T]\right\} \tag{1}$$

where E_t is the conditional expectation operator, conditional on information available at time t. $S(t) = [S_1(t), \ldots, S_m(t)]$ is a finite-m vector set of state variables which together with the consumer's current wealth $W(t)$ is sufficient to describe the state of the economy at time t. $c(t)$ denotes the instantaneous consumption flow selected at time t. U is a strictly concave, statedependent utility function for consumption and B represents utility from bequests at date T.

The evolution of the state variables S is described by a Markov system of Itô stochastic differential equations

$$\mathrm{d}S_i(t) = G_i(S,t)\mathrm{d}t + H_i(S,t)\mathrm{d}q_i, i = 1, 2, \ldots, m \tag{2}$$

where $G_i(S, t)$ is the instantaneous expected change in $S_i(t)$ per unit time at time t, H_i^2 is the instantaneous variance of the change in $S_i(t)$, where it is understood that these statistics are conditional on $S(t) = S$. The $\mathrm{d}q_i$ are Wiener processes with the instantaneous correlation coefficient per unit time between $\mathrm{d}q_i$ and $\mathrm{d}q_j$ given by the function $\eta_{ij}(S, t)$, $i, j = 1, \ldots, m$. At each point in time, the consumer chooses a consumption flow and allocates his wealth among n risky securities and a riskless security whose instantaneous rate of return per unit time is the interest rate $r(t)$. The rate of return dynamics on risky security j can be written as

$$\mathrm{d}P_j/P_j = \alpha_j(S,t)\mathrm{d}t + \sigma_j(S,t)\mathrm{d}z_j, \quad j = 1, 2, \ldots, n \tag{3}$$

where α_j is the instantaneous conditional expected rate of return per unit time; σ_j^2 is the conditional variance per unit time; and $\mathrm{d}z_j$ is a Wiener process. Denote by $\rho_{jk}(S, t)$ the instantaneous correlation coefficient per unit time between $\mathrm{d}z_j$ and $\mathrm{d}z_k$, and denote by $\mu_{ij}(S, t)$ the instantaneous correlation coefficient between $\mathrm{d}q_i$ and $\mathrm{d}z_j$, $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$.

The accumulation equation for the consumer's wealth can be written as

$$\mathrm{d}W = [rW + y - c]\mathrm{d}t + \sum_{j=1}^{n} w_j W\left[\mathrm{d}P_j/P_j - r\ \mathrm{d}t\right] \tag{4}$$

where $y = y(S, t)$ is the consumer's wage income; w_j is the fraction of his wealth allocated to risky security j at time t, and $\left[1 - \sum_j^n w_j\right]$ is the fraction allocated to the riskless asset.

The optimal consumption and portfolio rules, $c^*(W, S, t)$ and $w^*(W, S, t)$, are derived by the technique of stochastic dynamic programming. Among the first-order conditions to be satisfied by these optimal rules are the n conditions for the optimal portfolio holdings at time t, which can be expressed as $j = 1, 2, \ldots, n$

$$0 = \left\{\alpha_j - r - \left(\sum_1^n w_i^*\sigma_i\sigma_j\rho_{ij}W + \sum_1^m A_i\sigma_j H_i\mu_{ij}\right)\Big/K\right\} \times \frac{\partial U}{\partial c}\frac{\partial c^*}{\partial W}W \tag{5}$$

where

$$K \equiv -\partial U/\partial c/\left[\partial^2 U/\partial c^2 \cdot \partial c^*/\partial W\right]$$

and

$$A_i \equiv -\left[\partial c^*/\partial S_i + (\partial^2 U/\partial c \partial S_i)/(\partial^2 U/\partial c^2)\right]/(\partial c^*/\partial W), \quad i = 1, 2, \ldots, m.$$

By inspection, the manifest characteristic of the system of Eq. 5 is that it is linear in the optimal demands for risky assets. Therefore, if none of the risky assets is redundant, then standard matrix inversion can be used to solve Eq. 5 explicitly for these demands. That is,

$$w_j^*(t)W(t) = K\sum_1^n v_{kj}(\alpha_\kappa - r) + \sum_1^m A_i \zeta_{ij}, \quad j = 1, 2 \ldots, , , n \tag{6}$$

where v_{kj} is the $k-j$th element of the inverse of the variance– covariance matrix of returns

$$\left[\sigma_\iota \sigma_j \rho_{ij}\right] \text{ and } \zeta_{ij} \equiv \sum_1^n v_{kj}\sigma_k H_i \mu_{ik}.$$

By inspection, K, A_1,. . ., A_m are the only elements in Eq. 6 that depend on the individual investor's preferences or endowment. As an immediate consequence, it follows that there exist $(m + 2)$ portfolios ('mutual funds') constructed from linear combinations of the available securities such that, independent of preferences, wealth distribution, or planning horizon, all investors will be indifferent between choosing their portfolios from combinations of just these $(m + 2)$ funds or combinations of all n risky securities and the riskless security. This portfolio-separation theorem is, of course, vacuous if $m \geq n + 1$. If, however, $m \ll n$, then it implies a non-trivial reduction in the number of securities required to generate the set of optimal portfolios.

Although not unique, a set of funds which meets the criterion of the theorem is: fund no. 1 holds the riskless asset; fund no. 2 holds fraction $\sum_1^n v_{ki}(\alpha_k - r)$ in security j, $j = 1$,. . ., n and the balance in the riskless asset; for $i = 1, 2$,. . ., n, fund no.$(2 + i)$ holds fraction ζij in security j, $j = 1$,. . ., n and the balance in the riskless asset. Funds nos. 1 and 2, together generate the set of portfolios with maximum expected return for a given variance of the return (i.e. the mean-variance efficient set). Fund no. $(2 + i)$ provides the maximum feasible correlation between its return and the stochastic component of the instantaneous change in state variable $S_i(t)$, $i = 1, \ldots, m$. As discussed in detail in the cited Breeden and Merton papers, these latter portfolios serve the function of providing the best feasible hedges against utility losses caused by unanticipated changes in the state variables of the economy.

In the important case where the set of available securities is such that the return on fund no. $(2 + i)$ is perfectly correlated with the change in state variable $S_i(t)$ for each i, $i = 1, \ldots, m$, Breeden (1979) shows that the resulting intertemporal equilibrium allocations are Paretoefficient. This is also the condition under which it is possible to replicate the payoff structure for the complete set of pure Arrow–Debreu securities using continuous-trading dynamic portfolio strategies with a finite number of securities.

The dynamic strategies for replicating the payoffs to pure Arrow–Debreu securities can be derived in a similar fashion to the derivation of contingent-claim prices in Merton (1977). Suppose that among the available traded securities, portfolios can be constructed whose returns are instantaneously perfectly correlated with changes in each of the state variables, $[S_1(t), \ldots, S_m(t)]$. Without loss of generality, assume that these portfolios are the first m risky securities (i.e. $\mathrm{d}z_i = \mathrm{d}q_i$, $i = 1, 2. . .$, m).

Let $F(S, t)$ satisfy the linear partial differential equation

$$0 = \frac{1}{2}\sum_1^m \sum_1^m H_i H_j \eta_{ij} \frac{\partial^2 F}{\partial S_i \partial S_j} + \sum_1^m \left[G_j - H_j(\alpha_j - r)/\sigma_j\right] \frac{\partial F}{\partial S_j} + \frac{\partial F}{\partial t} - rF \tag{7}$$

subject to the boundary conditions: $0 \leq F(S,t) < \infty$ for all S and $t < \tau$; $F(S, \tau) = \delta\left[\overline{S}_1 - S_1(\tau)\right] \ldots \delta\left[\overline{S}_m - S_m(\tau)\right]$, where $\delta[]$ is the Dirac delta function and $\overline{S}_k$ are given parameters, $k = 1$,. . ., m. Under mid regularity conditions on the functions and r, a solution to Eq. 7 exists and is unique.

Consider the continuous-trading portfolio strategy which allocates fraction $x_j(t) = (\partial F/\partial S_j) H_j/[\sigma_j V(t)]$ to security j, $j = 1,\ldots, m$ and $\left[1 - \sum_1^m x_j(t)\right]$ to the riskless security at time t, where $V(t)$ denotes the value of the portfolio. It follows from Eq. 3 and the prescribed allocation that the dynamics of the portfolio value can be written as

$$\begin{aligned} dV &= V\left\{\left[\sum_1^m x_j(\alpha_j - r) + r\right]dt + \sum_l^m x_j\sigma_j dz_j\right\} \\ &= \left[\sum_1^m \frac{\partial F}{\partial S_j} H_j(\alpha_j - r)/\sigma_j + rV\right]dt + \sum_1^m \frac{\partial F}{\partial S_j} H_j dq_j \end{aligned} \tag{8}$$

because $dz_j = dq_j; j = 1, 2,\ldots, m$.

As a solution to Eq. 7, F is twice-continuously differentiable. Thus, Itô's Lemma can be used to describe the stochastic process for F as

$$\begin{aligned} dF = &\left(\frac{1}{2}\sum_1^m\sum_1^m H_i H_j \eta_{ij} \frac{\partial^2 F}{\partial S_i \partial S_j} + \sum_1^m \left[G_j \frac{\partial F}{\partial S_j}\right] + \frac{\partial F}{\partial S_j}\right)dt \\ &+ \sum_1^m \frac{\partial F}{\partial S_j} H_j dq_j \end{aligned} \tag{9}$$

where F is evaluated at $S = S(t)$ at each time t. Because F satisfies Eq. 7, Eq. 9 can be rewritten as

$$\begin{aligned} dF = &\left[\sum_1^m \frac{\partial F}{\partial S_j} H_j(\alpha_j - r)/\sigma_j + rF\right]dt \\ &+ \sum_1^m \frac{\partial F}{\partial S_j} H_j dq_j. \end{aligned} \tag{10}$$

From Eqs. 8 and 10, $dF = dV = r(F - V)dt$, which is an ordinary differential equation with solution $F[S(t), t] - V(t) = [F(S(0), 0) - V(0)] \exp\left[\int_0^t r(u)du\right]$. If, therefore, the initial investment in the portfolio is chosen so that $V(0) = F[S(0), 0]$ then $V(t) = F[S(t), t]$ for $0 \le t \le \tau$.

Thus, a dynamic portfolio strategy using $(m + 1)$ available securities has been constructed that has a payoff at $t = \tau$ of $\delta\left[\overline{S}_l - S_1(\tau)\right] \ldots \delta\left[\overline{S}_m - S_m(\tau)\right]$. By inspection of this payoff structure, it is evident that this security is the natural generalization of Arrow–Debreu pure state securities to an environment where there is a continuum of states defined by $\overline{S}_k$ and τ. By changing the time and state parameters τ and $\overline{S}_k$, one can generate all of the uncountable number of pure securities. Moreover, F, the solution to Eq. 7 used to implement each strategy, will also be the equilibrium price for the corresponding pure Arrow–Debreu security.

Continuous trading, like any other continuous-revision process, is of course an abstraction from physical reality. If, however, the length of time between revisions is very short, then the continuous-trading optimal solutions will be a reasonable approximation to their discrete-time counterparts (see Samuelson 1970 and Merton 1975b, 1982b). From the work of Magill and Constantinides (1976), this conclusion appears to be robust even in the presence of transactions costs, which cause trading to be discrete almost certainly.

Whether the length of time between revisions is short enough for the continuous solution to provide a good approximation must be decided on a case-by-case basis by making a relative comparison with other time scales in the problem. The continuous-trading assumption appears to be especially appropriate for the analysis of security markets where the aggregate trading volume is large, the minimum unit-size for a transaction is relatively small, and the length of calendar time between successive transactions is quite short.

The continuous analysis may also provide a valid approximation in problems where the calendar length of time between revisions is not short. For example, Bourguignon (1974), Bismut (1975), and Merton (1975a) use this mode of analysis to extend the Solow model of economic growth to an uncertain environment and to analyse the stochastic Ramsey problem. It is the practice in such models to neglect 'short-run' business cycle fluctuations and to assume full employment. Moreover, the exogenous factors usually assumed to affect the time path of the economy in these models are either demographic or technological changes. Since major changes in either factor typically take rather long periods of time, the

length of time between revisions in the capital stock, although hardly instantaneous, may well be quite short, relative to the time scale of the exogenous processes.

See Also

- ▶ Continuous-Time Stochastic Processes
- ▶ Finance
- ▶ Options

Bibliography

Bismut, J.M. 1975. Growth and optimal intertemporal allocation of risks. *Journal of Economic Theory* 10: 239–257.

Black, F., and M. Scholes. 1973. The pricing of options and corporate liabilities. *Journal of Political Economy* 81: 637–654.

Bourguignon, F. 1974. A particular class of continuous-time stochastic growth models. *Journal of Economic Theory* 9: 141–158.

Breeden, D.T. 1979. An intertemporal asset pricing model with stochastic consumption and investment opportunities. *Journal of Financial Economics* 7: 265–296.

Cox, J.C., J.E. Ingersoll Jr., and S.A. Ross. 1985a. A theory of the term structure of interest rates. *Econometrica* 53: 385–408.

Cox, J.C., J.E. Ingersoll Jr., and S.A. Ross. 1985b. An intertemporal general equilibrium model of asset prices. *Econometrica* 53: 363–384.

Duffie, D., and C. Huang. 1985. Implementing Arrow–Debreu equilibria by continuous trading of a few long-lived securities. *Econometrica* 35: 1337–1356.

Huang, C. 1985. Information structure and equilibrium asset prices. *Journal of Economic Theory* 35: 33–71.

Magill, M.J.P., and G.M. Constantinides. 1976. Portfolio selection with transactions costs. *Journal of Economic Theory* 13: 245–263.

Malliaris, A.G., and W.A. Brock. 1982. *Stochastic methods in economics and finance*. Amsterdam: North-Holland.

Merton, R.C. 1969. Lifetime portfolio selection under uncertainty: The continuous-time case. *Review of Economics and Statistics* 51: 247–257.

Merton, R.C. 1971. Optimum consumption and portfolio rules in a continuous-time model. *Journal of Economic Theory* 3: 373–413.

Merton, R.C. 1973. An intertemporal capital asset pricing model. *Econometrica* 41: 867–887.

Merton, R.C. 1975a. An asymptotic theory of growth under uncertainty. *Review of Economic Studies* 42: 375–393.

Merton, R.C. 1975b. Theory of finance from the perspective of continuous time. *Journal of Financial and Quantitative Analysis* 10: 659–674.

Merton, R.C. 1977. On the pricing of contingent claims and the Modigliani-Miller theorem. *Journal of Financial Economics* 5: 241–249.

Merton, R.C. 1982a. On the microeconomic theory of investment under uncertainty. In *Handbook of mathematical economics*, vol. 2, ed. K.J. Arrow and M.D. Intriligator. Amsterdam: North-Holland.

Merton, R.C. 1982b. On the mathematics and economics assumptions of continuous-time models. In *Financial economics: Essays in honor of Paul Cootner*, ed. W.F. Sharpe and C.M. Cootner. Englewood Cliffs: Prentice-Hall.

Radner, R. 1972. Existence of plants, prices, and price expectations in a sequence of markets. *Econometrica* 40: 289–303.

Samuelson, P.A. 1970. The fundamental approximation theorem of portfolio analysis in terms of means, variances, and higher moments. *Review of Economic Studies* 32: 537–542.

Continuous-Time Stochastic Processes

Chi-Fu Huang

Applications of continuous-time stochastic processes to economic modelling are largely focused on the areas of capital theory and financial markets. In these applications as in mathematics generally, the most widely studied continuous time process is a Brownian motion – so named for its early application as a model of the seemingly random movements of particles which were first observed by the English botanist Robert Brown in the 19th century. Einstein (1905), in the context of statistical mechanics, is generally given credit for the first mathematical formulation of a Brownian motion process. However, an earlier development of an equivalent continuous-time process is provided by Louis Bachelier (1900) in his theory of stock option pricing. Framed as an abstract mathematical process, a Brownian motion {B(t); t ∈ R+} is described by the following properties: (1) for $0 : (s \leq t < \infty, B(t) - B(s)$ is a normally distributed random variable with mean zero and variance $t - s$; (2) for $0 \leq t_0 < t_1 < \cdots < t_1 < \infty$,

$$\{B(t_0); B(t_k) - B(t_{k-1}), k = 1, \ldots, l\}$$

is a set of independent random variables.

From this construction, Doob, Feller, Itô, Wiener, among others went on to develop the general theory of continuous-time stochastic processes.

During the half century of this development of the theory, its application in economics was confined primarily to the formulation and testing of hypotheses concerning time series properties of economic variables. It was not until the 1950s and early 1960s that the theory of continuous-time stochastic processes found its way into economic theory. Motivated by the rediscovery of Bachelier's work on options by L.J. Savage, Samuelson (1965) presents a theory of rational warrant pricing. Unlike Bachelier's assumption of a Brownian motion for a stock price process, Samuelson posits that the *logarithm* of a stock price follows a Brownian motion, and thereby, ensures that model stock prices exhibit non-negativity as required by limited liability. This process, called a *geometric Brownian motion* by Samuelson, remains to this day the prototypical process used by economists to describe stock price behaviour. Working with Samuelson, McKean (1965) uses the theory of optimal stopping to provide a rigorous derivation of the warrant price in Samuelson's theory.

Although it is the standard mode of analysis for warrant and option pricing theory today, the celebrated work on the stochastic integration by K. Itô (1944, 1951) was not introduced into economic analysis until the late 1960s. Merton (1969, 1971) was the first to use Itô's stochastic calculus in economics. He analysed an agent's optimal consumption and portfolio policies in a continuous time economy where asset prices are Itô processes.

Itô's contribution to the theory of stochastic processes lies in a definition of an integral with desired properties when the integrator is a Brownian motion. A pathwise definition in the Stieltjes sense may fail since a Brownian motion has sample paths that are nowhere differentiable with probability one. For a classical treatment of the Itô integral, see also Itô and McKean (1965) and McKean (1969). A good reference for modern treatments can be found in Chung and Williams (1983).

Itô's definition of a stochastic integral, in contrast to that of Stratonovich (1966), is much better suited for analysing intertemporal economic decision making. The *non-anticipating* integrand in Itô's definition captures the economic constraint that agents cannot anticipate future speculative price movements.

The most useful result of Itô's stochastic calculus is the so-called Itô's lemma: any twice continuously-differentiable function of an Itô process is itself an Itô process. This implies that the agent's wealth process is an Itô process and therefore the Bellman equation in the stochastic dynamic programming problem becomes a second-order partial differential equation. The latter allows one to analyse the portfolio problem by looking at just the first two moments of price processes and to achieve sharp characterizations. Merton (1973a) applied this technique further to study equilibrium relations among risky asset prices and arrived at the *Intertemporal Capital Asset Pricing Model*.

The introduction of Itô's stochastic calculus opened a whole new world for economists. With it, most of the static utility maximization models are readily extended to a dynamic setting with uncertainty. The continuous time set-up allows one to work with differential equations rather than with difference equations. For applications to capital theory and economic growth, see Bismut (1975), Brock and Magill (1979), and Merton (1975); to asset pricing models, see Cox et al. (1985a, b).

Itô's work was later extended by Kunita and Watanabe (1967) to the case where integrators are square-integrable martingales. They also proved a *martingale representation theorem* for a Brownian motion: any square-integrable martingale adapted to a Brownian motion filtration (see below) is representable as an Itô integral.

The most general notions of a stochastic process and a stochastic integral to date are in the terrain of the so-called *French School Probability Theory*, or the *General Theory of Processes*. Very abstract, and surely developed for intrinsic intellectual reasons, it nevertheless seems to have been

invented for the study of financial markets; for references, see Dellacherie and Meyer (1978, 1982), Jacod (1979), and Meyer (1966, 1976).

Although making no explicit use of French probability theory, the seminal paper of Black and Scholes (1973) in the pricing of stock options nevertheless opens up the possibility of its application in financial economics. This work was subsequently generalized and formalized by Merton (1973b, 1977). The idea is that the payoff of a stock option can be *replicated* by continuous trading in its underlying stock and a riskless asset. The replicating strategy is *self-financing* in that after the beginning of this strategy there are neither additional funds invested into it nor funds withdrawn out of it. Thus, to rule out arbitrage opportunities, the stock option must sell for the exact value of the replicating portfolio at any point in time.

Black and Scholes's theory provided a strong incentive for financial economists to study continuous time stochastic processes. The key observation in this literature was made by Cox and Ross (1976). They noted that since the expected rate of return of the stock does not enter the Black and Scholes pricing formula for a stock option, the price of an option must be determined as if investors were risk neutral and had probability beliefs such that the stock earns an expected rate of return equal to the riskless rate. Harrison and Kreps (1979) formalized this observation in showing that *any* arbitrage-free price system can be converted into a martingale through a change of an equivalent probability after a suitable normalization. Note that this martingale connection of an arbitrage-free price system was vaguely foreshadowed in Samuelson and Merton (1969).

Harrison and Kreps (1979) and Harrison and Pliska (1981) make clear that the answer to whether a contingent claim can be replicated by dynamic trading is intimately related to the martingale representation theorem. A sketch of their arguments will be given.

Taken as primitive is a complete separable probability space $(\Omega, \mathcal{F}, P)$ and a *filtration* $\mathbf{F} = \{\mathcal{F}_t; t \in [0, 1]\}$. A filtration is an increasing family of sub-sigma-algebras of $\mathcal{F}$ representing information revelation over time. For simplicity, we take the time span of the economy to be [0,1]. We assume that agents are endowed with the same information structure **F**. Readers can think of a filtration to be like an event tree in a discrete time finite state setting. We also assume that agents at time zero knows that the true state of the nature is an element $\omega \in \Omega$, which they will learn at time one.

Agents can only consume at time one. For simplicity again, we take the commodity space to be the space of square-integrable random variables defined on $(\Omega, \mathcal{F}, P)$, denoted by $L^2(P)$.

There are $N + 1$ long-lived securities traded indexed by $n = 0, 1, \ldots, N$. A long-lived security is a security available for trading all the time in [0,1] and is represented by a price process. $\{S_n(t)\}$ It pays a dividend only at time one and is equal to $S_n(1)$ almost surely. Price processes are semimartingales (adapted to **F**) and $S_n(1) \in L^2(P)$. In modelling a dynamic asset trading economy, before anything interesting can be said, one has to formulate a budget constraint. That naturally involves stochastic integrals. Jacod (1979) has shown that for stochastic integrals to have desired properties, it is necessary that integrators be semimartingales. Thus, semimartingale price processes can be assumed without loss of generality.

In a Walrasian economy, only relative prices are determined. Thus we can assume that the price system has been normalized such that $S_0(t) = 1$ $\forall t \in [0, 1]$. We will call the 0th security the riskless security and the rest risky securities.

A trading strategy is an $(N + 1)$-dimensional predictable process $\theta = \{\theta_n(t); n = 0, 1, -N\}$, where we interpret $\theta_n(t)$ to be the number of shares of security n held from $t -$ to t before trading at time t. A process is predictable if its values at time t depend only upon the information available strictly before time t. Given the interpretation of θ, predictability is a natural information constraint.

A trading strategy is said to be *simple* if it is bounded and changes its value at most at a finite number of time points in [0,1].

A trading strategy θ; is said to be *self-financing* if the stochastic integral

$$\int_0^t \theta(s)^T ds(s)$$

is well-defined and if

$$\theta(t)^T\, S(t) = \theta(0)^T S(0) + \int_0^t \theta(s)^T \mathrm{d}S(s) \quad \forall t \in [0,\ 1]\ \text{a.s.}, \tag{1}$$

where ^{T} denotes transpose. That is, the value of the portfolio θ at time t is equal to its initial value plus accumulated capital gains or losses from time zero to time t. There are neither new investments into nor withdrawals of funds out of the portfolio. This is just a natural budget constraint.

Harrison and Kreps (1979) and Kreps (1981) show that if all the simple of self-financing trading strategies are allowed and if arbitrage opportunities are absent, then there exists a probability measure Q equivalent to P such that the Radon-Nikodym derivative $\xi \equiv \mathrm{d}Q/\mathrm{d}P$ is an element of $L^2(P)$ and that S is a martingale under Q, or a Q-martingale. Fix Q and note that since P and Q are equivalent, all the *a.s.* statements to follow apply to both.

Now we can specify the space of admissible strategies $\Theta\ [S]$. A self-financing trading strategy θ is admissible if

$$\int_0^t \theta(s)^T \mathrm{d}S(s) \quad t \in [0, 1]$$

is a Q-martingale and $\theta\ (1)^T\ S(1) \in L^2(P)$. [See Jacod (1979) for sufficient conditions for this to be true.] Then one can show that given $\Theta\ [S]$ indeed there are no arbitrage opportunities.

A contingent claim is an element of $L^2(P)$. A contingent claim x is said to be *marketed* if it can be dynamically manufactured by an admissible trading strategy. Formally, x is marketed if there exists $\theta \in \Theta\ [S]$ such that

$$x = \theta(0)^T\ S(0) + \int_0^1 \theta(t)^T \mathrm{d}S(t) \quad a.s.$$

The value of x at time t is $\theta\ (t)^T\ S(t)$. By the definition of admissibility, we have

$$\int_0^t \theta(s)^T \mathrm{d}S = \sum_{n=1}^{N} \int_0^t \theta_n(s)\ \mathrm{d}S_n\ (s). \tag{2}$$

Now here is the key observation. Let x be a contingent claim. We know from relation (2) that if it is marketed, its value over time is equal to its initial value at time zero plus a stochastic integral with respect to N Q-martingales. Conversely, a contingent claim x marketed if the conditional expectation $E^*\ [x \mid \mathcal{F}_t]$, which is a Q-martingale, can be represented by a stochastic integral with respect to the N Q-martingales $\{S_n(t);\ n = 1, 2, \ldots, N\}$. [Here we should remark that any $x \in L^2(P)$ has a finite expectation under Q by the Cauchy–Schwarz inequality.] This observation turns on the machinery of the martingale representation theorem in the study of market completeness.

The security markets are said to be *dynamically complete* if all contingent claims are marketed. From the above discussion, it follows that markets are dynamically complete if all Q-martingales are representable as stochastic integrals with respect to the N risky Q-martingale prices. In such event, the N Q-martingales are said to have the *martingale representation property*. Readers might be curious by now that the riskless asset seems to disappear from the story. Indeed, whether a contingent claim is generated by a (not necessarily self-financing) trading strategy does not depend upon the riskless asset after time zero. The riskless asset, however, is a vehicle through which the budget is balanced over time.

The contribution made by Harrison, Kreps, and Pliska is methodological. They make available a powerful machinery for the study of financial/capital markets: the theory of martingales. Now we shall present some consequences of their work.

Since Merton's (1969, 1971, 1973a) analyses of optimal intertemporal consumption-portfolio policies and their implications on equilibrium asset prices, the conditions under which a price system is representable is an Itô process had been an open question for more than a decade. A short answer found in Huang (1985a) is as follows: Take the set-up of the economy as above and

C

assume henceforth that there are no arbitrage opportunities. Moreover, assume that the information structure **F** is a Brownian motion filtration. We know S is a Q-martingale, so we can write

$$S_n(t) = E^*\left[S_n(1)\middle|\mathcal{F}_t\right] \quad a.s.$$
$$= \frac{E\left[S_n(1)\xi\middle|\mathcal{F}_t\right]}{E\left[\xi\middle|\mathcal{F}_t\right]} \quad a.s.,$$

where the second equality follows from the Bayes' rule, and where we recall that $\xi = \mathrm{d}Q/\mathrm{d}P$, which is strictly positive by the fact that Q and P are equivalent. The numerator and the denominator of the above relation are both P-square integrable martingales. By the martingale representation theorem of Kunita and Watanabe (1967), we know that any P-square integrable martingale is representable as an Itô integral. Then S_n is an Itô process by Itô lemma. Hence any arbitrage-free price system is an Itô process when the information structure is a Brownian filtration.

We can also study the sample path properties of a price system, which relates to examining empirically the so-called *efficient market hypothesis*. Much of the empirical work in financial economics and accounting concerns the response of capital/financial asset prices to information. The null hypothesis in this work is typically that the capital/financial markets are *efficient* in the sense that prices rapidly adjust to *new* information. But is it true that prices only make large adjustments at surprises and what exactly is a *surprise*, mathematically? Here we turn to the classification of stopping times in the general theory of processes. In this context, a surprise is a non-predictable stopping time. We also know that a martingale must be continuous at predictable stopping times (provided that a minor technical condition is satisfied). Thus, S can make discrete changes only at nonpredictable stopping times or at surprises. This and other related issues can be found in Huang (1985a, b). A reference for the classification of stopping times is Dellacherie and Meyer (1978).

So we discover that a price system must be Itô process when the information is a Brownian motion filtration and when there are no arbitrage opportunities. There still remain further questions: does there exist an equilibrium where equilibrium price processes are Itô processes? More importantly, does there exist an equilibrium where although there are only a finite number of long-lived securities traded, the markets are dynamically complete and thus the equilibrium allocation is Pareto optimal? Note that in the Arrow–Debreu equilibrium theory, markets for all contingent claims are available at time zero. Agents trade to a Pareto optimal allocation. There is no need and no incentive for the markets to reopen after time zero. Of course, this does not conform with actual market structures. We do not have a complete set of contingent markets. What we do have are constantly-open financial markets where a finite number of long-lived assets are traded. Thus it is important to know whether there exists an equilibrium in such a world and to know the efficiency of the resulting allocation.

It follows from the earlier discussion on the martingale representation property of risky price processes that what is needed for an affirmative answer to the above questions: is that there be a riskless security with unit price throughout and a finite number of risky long-lived securities that have the martingale representation property. What complicates the story, however, is that the demand and supply of the long-lived securities must be equal in equilibrium. Thus those securities must be picked carefully. Moreover, it is not true that a finite number of martingales having the martingale representation property can always be found. Duffie and Huang (1985, 1986b) and Duffie (1986a), in exchange as well as production economies, demonstrated a procedure to select long-lived securities having the desired properties and conditions under which the number is finite.

The martingale connection of an arbitrage-free system has been generalized to economies where securities can pay dividends and agents can consume at any time in [0,1]. After a suitable normalization, a price system plus the accumulated dividends form a martingale under an equivalent probability measure. This is done in Huang (1985b). Similar theory is also valid in economies where agents have differential information. Interested readers are referred to Duffie and Huang (1986b) for details.

Although the focus of research has been on capital theory and financial markets, applications of the theory of continuous time stochastic processes to economic problems outside these areas can be found. For example, Duffie (1986b) applies classical potential theory as in the context of Markov processes to valuation of securities, and Li (1984) examines the stochastic theory of the firm in continuous time. He uses point processes to model stochastic demands for commodities and endogenizes a firm's demand for inventories, among other things. For applications of the theory of optimal stopping to game theory, see Hugues (1974) for zero-sum stopping games, and Huang and Li (1986) for nonzero-sum stopping games.

See Also

- ▶ Capital Asset Pricing Model
- ▶ Continuous-Time Stochastic Models
- ▶ Options

Bibliography

Bachelier, L. 1900. *Théorie de la speculation*. Paris: Gauthier-Villars.

Bismut, J. 1975. Growth and optimal intertemporal allocation of risks. *Journal of Economic Theory* 10: 239–257.

Black, F., and M. Scholes. 1973. The pricing of options and corporate liabilities. *Journal of Political Economy* 81: 637–654.

Brock, W., and M. Magill. 1979. Dynamics under uncertainty. *Econometrica* 47: 843–868.

Chung, K., and R. Williams. 1983. *An introduction to stochastic integration*. Boston: Birkhauser.

Cox, J., and S. Ross. 1976. The valuation of options for alternative stochastic processes. *Journal of Financial Economics* 3: 145–166.

Cox, J., J. Ingersoll, and S. Ross. 1985a. An intertemporal general equilibrium model of asset prices. *Econometrica* 53: 363–384.

Cox, J., J. Ingersoll, and S. Ross. 1985b. A theory of the term structure of interest rates. *Econometrica* 53: 385–408.

Dellacherie, C., and P. Meyer. 1978. *Probabilities and potential A: General theory of process*. New York: North-Holland.

Duffie, D. 1986a. Stochastic equilibria: Existence, spanning number, and the 'no expected gains for trade' hypothesis. *Econometrica*.

Duffie, D. 1986b. Price operators: Extensions, potentials, and the Markov valuation of securities. Research Paper No. 813. Graduate School of Business, Stanford University.

Duffie, D., and C. Huang. 1985. Implementing Arrow–Debreu equilibria by continuous trading of few long-lived securities. *Econometrica* 53: 1337–1356.

Duffie, D., and C. Huang. 1986a. Multiperiod securities markets with differential information: Martingales and resolution times. *Journal of Mathematical Economics* 15: 283.

Duffie, D., and C. Huang. 1986b. *Stochastic production-exchange equilibria*. Stanford: Graduate School of Business, Stanford University.

Duffie, C., and P. Meyer. 1982. *Probabilities and potential B: Theory of martingales*. New York: North-Holland.

Einstein, A. 1905. On the movement of small particles suspended in a stationary liquid demanded by the molecular kinetic theory of heat. *Annals of Physics* 17: 549.

Harrison, M., and D. Kreps. 1979. Martingales and arbitrage in multiperiod securities markets. *Journal of Economic Theory* 20: 381–408.

Harrison, M., and S. Pliska. 1981. Martingales and stochastic integrals in the theory of continuous trading. *Stochastic Processes and their Applications* 11: 215–260.

Huang, C. 1985a. Information structure and equilibrium asset prices. *Journal of Economic Theory* 35: 33–71.

Huang, C. 1985b. Information structure and viable price systems. *Journal of Mathematical Economics* 14: 215–240.

Huang, C., and L. Li. 1986. Continuous time stopping games, Working Paper No. 1796–86, Sloan School of Management, MIT.

Hugues, C. 1974. Markov games. Technical Report No.33, Department of Operations Research, Stanford University.

Itô, K. 1944. Stochastic integrals. *Proceedings of the Imperial Academy* 22: 519–524. Tokyo.

Itô, K. 1951. *On stochastic differential equations*, Memoirs of the American Mathematical Society. Rhode Island: The American Mathematical Society.

Itô, K., and H. McKean. 1965. *Diffusion processes and their sample paths*. New York: Springer.

Jacod, J. 1979. *Calcul stochastique et problemes de martingales*, Lecture Notes in Mathematics, vol. 714. New York: Springer.

Kreps, D. 1981. Arbitrage and equilibrium in economies with infinitely many commodities. *Journal of Mathematical Economics* 8: 15–35.

Kunita, H., and S. Watanabe. 1967. On square-integrable martingales. *Nagoya Mathematics Journal* 30: 209–245.

Li, L. 1984. A stochastic theory of the firm. Unpublished PhD thesis, Northwestern University.

McKean, H. 1965. Appendix: A free boundary problem for the heat equation arising from a problem in mathematical economics. *Industrial Management Review* 6: 32–39.

McKean, H. 1969. *Stochastic integrals*. New York: Academic.
Merton, R. 1969. Lifetime portfolio selection under uncertainty: The continuous case. *Review of Economic and Statistics* 51: 247–257.
Merton, R. 1971. Optimum consumption and portfolio rules in a continuous time model. *Journal of Economic Theory* 3: 373–413.
Merton, R. 1973a. An intertemporal capital asset pricing model. *Econometrica* 41: 867–888.
Merton, R. 1973b. Theory of rational option pricing. *Bell Journal of Economics and Management Science* 4: 141–183.
Merton, R. 1975. An asymptotic theory of growth under uncertainty. *Review of Economic Studies* 42: 375–393.
Merton, R. 1977. On the pricing of contingent claims and the Modigliani–Miller theorem. *Journal of Financial Economics* 5: 241–249.
Meyer, P. 1966. *Probability and potentials*. Waltham: Blaisdell Publishing Company.
Meyer, P. 1976. Un cours sur les integrales stochastiques. In *Seminaires de Probabilité X.* Lecture Notes in Mathematics 511. New York: Springer.
Samuelson, P. 1965. Rational theory of warrant pricing. *Industrial Management Review* 6: 13–32.
Samuelson, P., and R. Merton. 1969. A complete model of warrant pricing that maximizes utility. *Industrial Management Review* 10: 17–46.
Stratonovich, R. 1966. A new representation for stochastic integrals and equations. *SIAM Journal of Control* 4: 362–371.

Contract Theory

David Martimort

Abstract

This article offers a brief overview of contract. It focuses on the theory of complete contracts and the three associated paradigms of adverse selection, moral hazard and non-verifiability. By showing difficulties in allocating resources between asymmetrically informed partners, contract theory has deeply changed our view of the functioning of organizations and markets.

Keywords

Adverse selection; Asymmetrical information; Bayesian-Nash equilibrium; Collusion; Contract theory; Cost observability; Free-rider problem; Incentive compatibility; Incentive constraints; Incomplete contracts; Informativeness principle; Insurance; Laffont, J.-J.; Limited liability; Monotonicity; Moral hazard; Multi-agent organizations; Non-verifiability; Optimal contract; Pontryagyn principle of optimality; Principal and agent; Revelation principle; Risk aversion; Risk neutrality; Sharecropping; Spence–Mirrlees condition; Tournaments

JEL Classifications

D0

As with so many major concepts in economics, contract theory was introduced by Adam Smith who, in his monumental *Wealth of Nations* (1776, book III, ch. 2), considered the relationship between peasants and farmers through this lens. For instance, he pointed out the perverse incentives provided by sharecropping contracts, widespread in 18th-century Europe. However, it is fair to say that the issues of incentives and contract theory were largely ignored by economists until the end of the 20th century. By then, the focus of economic theory was on the working of markets and price formation. Firms were viewed only as production technologies, and the issue of the separation between ownership and control was most often put aside. This black-box approach was, of course, quite unsatisfactory. At the turn of the 1970s, with the methodological revolution of game theory, more emphasis was placed on strategic interactions between a small number of players in a world where informational problems matter. From this new perspective, the allocation of resources is no longer ruled by the price system but by *contracts* between asymmetrically informed partners. Contract theory has deeply changed our view of the functioning of organizations and markets.

This article aims to provide a brief overview of contract theory, stressing a few major insights and illustrating them with useful applications. Due to space constraints, it does not do justice to several aspects of contract theory, and will mostly reflect my own tastes in the field. In particular, I focus on the so-called *theory of complete contracts*, leaving

aside the burgeoning theory of incomplete contracts which is covered elsewhere in this dictionary. Successive sections deal respectively, with adverse selection, moral hazard and non-verifiability: the three different paradigms which have been used in the field of complete contract theory. Since the distinction between complete and incomplete contracts is easier to draw once these notions have already been explained, I will postpone such discussion to the end of the article.

Adverse Selection

Consider the following buyer–seller relationship as the archetypical example of contractual relationship between a principal (the buyer) and his agent (the seller) who produces some good or service on his behalf. The mere delegation of this task to the agent gives the agent access to private information about the technology. This *adverse selection* environment is captured by assuming that a technological parameter θ is known only by the agent. It is drawn from a distribution in an exogenous type space Θ which is common knowledge. Neither the principal nor a court of law observes this parameter. Contracts cannot specify outputs and prices as a function of the realized state of nature.

The buyer enjoys a net benefit $S(\theta, q) - t$ when buying q units of output at a price t. The seller enjoys a profit $t - C(\theta, q)$ from producing that good. We will assume that these functions are concave in q. Notice that the state of nature θ might affect both the agent's and the principal's utility functions. This can, for instance, be the case if this parameter also determines the quality of the good to be traded.

Under complete information, efficiency requires that the buyer and the seller trade the first-best quantity $q^*(\theta)$ such that the buyer's marginal benefit from consumption equals the seller's marginal cost of production:

$$\frac{\partial S}{\partial q}(\theta, q^*(\theta)) = \frac{\partial C}{\partial q}(\theta, q^*(\theta)). \tag{1}$$

Many mechanisms or institutions lead to this outcome. Both the price mechanism and a take-it-or-leave-it offer by one party to the other would achieve the same allocation, although with different distributions of the surplus between the traders. If the principal retains all bargaining power (for instance, because there is a competitive fringe of potential sellers), he could offer a forcing contract stipulating an output $q^*(\theta)$ and a transfer $t^*(\theta)$ which just covers the seller's cost. This forcing contract maximizes the buyer's net gains from trade and leaves the seller just indifferent between participating or not.

In what follows, we mostly focus on the case where the uninformed principal has full bargaining power in contracting. In this framework, the *contract* between the buyer and the seller does not only have the allocative and distributive roles it has under complete information. It also has the role of *communicating information* from the informed party to the uninformed party. This communication role suggests that the informed party should be given a choice among different options and that this choice should reveal information about the adverse selection parameter.

A first step in the analysis consists of describing the set of allocations which are feasible under asymmetric information. The basic tool for doing so is the revelation principle (see Gibbard 1973; Green and Laffont 1977; Dasgupta et al. 1979; Myerson 1979, among others), which states that there is no loss of generality in restricting the analysis to *revelation mechanisms* that are *direct*, that is, of the form $\left\{t\left(\hat{\theta}\right); q\left(\hat{\theta}\right)\right\}_{\hat{\theta} \in \Theta}$ with $\hat{\theta}$ a message ('report') sent by the informed seller to the uninformed buyer, and *truthful*, that is, such that the agent finds it optimal to report his true type.

Therefore, incentive feasible contracts satisfy the following *incentive* constraints

$$\begin{aligned} &t(\theta) - C(\theta, q(\theta)) \geq t\left(\hat{\theta}\right) \\ &- C\left(\theta, q\left(\hat{\theta}\right)\right) \forall \left(\theta, \hat{\theta}\right) \in \Theta^2. \end{aligned} \tag{2}$$

To be acceptable, a contract must also satisfy the seller's participation constraints

$$t(\theta) - C(\theta, q(\theta)) \geq 0 \quad \forall \theta \in \Theta \tag{3}$$

which ensure that, irrespective of his type, the agent by contracting gets at least his reservation payoff (exogenously normalized to zero).

Once the set of incentive feasible allocations is described, the analysis may proceed further. Keeping in mind that the uninformed buyer designs his offer under asymmetric information, we might characterize an *optimal contract*. Such a contract maximizes the uninformed buyer's expected net surplus subject to the feasibility constraints (2) and (3).

Much of the theoretical literature developed over the 1980s and early 1990s has investigated the structure of the set of incentive feasible allocations and its consequences for optimal contracting. A key property is the so-called Spence–Mirrlees condition (see Spence 1973, 1974; Mirrlees 1971) for early contributions which put forward that condition). This condition is satisfied when the slope of the agent's indifference curves can be ranked with respect to his type. In our example, this condition holds when $\frac{\partial^2 C}{\partial \theta \partial q} > 0$, that is, when higher types also have higher marginal costs and should thus produce less. Therefore, the monotonicity condition

$$q(\theta) \geq q(\theta') \quad \text{for } \theta \leq \theta' \tag{4}$$

is a direct consequence of the incentive constraints. The Spence–Mirrlees condition can be viewed as a regularity assumption making the incentive problem well-behaved. It ensures that only incentive constraints between 'nearby' types matter in the optimization. Intuitively, this means that the seller with a given marginal cost may be tempted to overstate slightly its costs, receiving the higher transfer targeted to less efficient types but producing at a lower marginal cost. By so doing, this more efficient type receives an information rent. Once these local constraints are taken into account and when the Spence–Mirrlees condition holds, the incentives to mimic more distant types are no longer relevant. With this reduction of the set of relevant incentive constraints, the principal's optimization problem is significantly simplified.

The result of this optimization is straightforward. Inducing information revelation by the most efficient types requires giving up an *information rent* to those types. The basic intuition of most adverse-selection models is that reducing this rent requires production to be distorted. For instance, when efficient types want to mimic less efficient ones, the latter's allocation should be made less attractive. This is obtained by distorting their production downward and modifying transfers accordingly.

To see more formally the nature of the output distortion, consider the case where types are distributed over a compact set $\left[\underline{\theta}, \overline{\theta}\right]$ according to the cumulative distribution function $F(\cdot)$ (with a positive density $f(\cdot)$). The second-best optimal output $q^{SB}(\theta)$ under adverse selection is the solution to:

$$\frac{\partial S}{\partial q}\left(\theta, q^{SB}(\theta)\right) = \frac{\partial C}{\partial q}\left(\theta, q^{SB}(\theta)\right) + \frac{F(\theta)\partial^2 C}{f(\theta)\partial q \partial \theta}\left(\theta, q^{SB}(\theta)\right). \tag{5}$$

Condition (5) states that, for any type θ, the buyer's marginal benefit must equal the seller's *marginal virtual cost* (see Laffont and Martimort 2002, chs 2 and 3, for details). The virtual cost of a given type takes into account not only its cost of production but also the cost of deterring other types (here more efficient types) from mimicking that type. The allocation is no longer efficient, as under complete information, but *interim efficient* in the sense of Holmström and Myerson (1983).

Condition (5) is crucial, and is found in various forms in any adverse-selection model. It states that, under asymmetric information, there is a fundamental trade-off between implementing allocations close to efficiency and giving information rents to the most efficient types to induce information revelation. This trade-off calls for distortions away from efficiency.

Provided that the output schedule defined by (5) satisfies the monotonicity condition (4), this is the exact solution of our problem. To guarantee monotonicity, on top of assumptions on the concavity of $S(\cdot)$ and $\frac{\partial S}{\partial \theta}(\cdot)$, convexity of $C(\cdot)$ and $\frac{\partial C}{\partial \theta}(\cdot)$, $\frac{\partial^2 C}{\partial \theta \partial q}(\cdot) > 0$, $\frac{\partial^3 C}{\partial \theta^2 \partial q}(\cdot) > 0$ and $\frac{\partial^2 S}{\partial \theta \partial q}(\cdot) < 0$,

one needs also to impose a property on the type distribution, the so-called *monotonicity of the hazard rate* $\frac{F(\theta)}{f(\theta)}$ (see Bagnoli and Bergstrom 2005). Otherwise, the optimal contract may entail some area of pooling such that all types belonging to a set with positive measure produce the same amount and are paid the same price. The optimal solution may then be obtained using 'ironing techniques' (see for instance Guesnerie and Laffont 1984).

Direct Extensions

Adverse-selection methodology has been successfully extended in various directions allowing for multidimensional types (Armstrong and Rochet 1999), and/or multiple outputs (Laffont and Tirole 1993, ch. 3), and type-dependent reservation utilities (Lewis and Sappington 1989; Jullien 2000). There, the analysis is substantially more complex as types can no longer be ranked as easily as in the model sketched above. The Spence–Mirrlees condition might fail to hold and global incentive constraints may bind, leading to pooling allocations being optimal. Another interesting extension is the case of hidden knowledge, in which contracting takes place before the agent becomes informed. The logic of such models is very close to that we discuss below in the section on moral hazard. In a nutshell, the trade-off between allocative efficiency and rent extraction is now replaced by the trade-off between insuring the agent against shocks on costs and inducing him to reveal his cost once it is known. Output distortions still arise (see Laffont and Martimort 2002, ch. 2, for details). Others have endogenized the asymmetric information structure and examined the incentives to learn about the unknown parameter (see, for instance, Crémer et al. 1998). Finally, there exists a literature that considers the case where the principal is the informed party (Maskin and Tirole 1990, 1992). New difficulties arise from the fact that the mere offer of the contract may signal information.

Multi-agent Organizations

The most important extensions of the adverse selection paradigm certainly concern multi-agent organizations. Such complex organizations emerge because of the need to share common resources, produce public goods, internalize production externalities or enjoy information economies of scale. Although any such reason calls for a specific analysis, a few common themes of the literature can be highlighted by remaining at a rather general level.

Regarding the implementation concept, different notions of incentive compatibility may be used depending on the context. First, agents may know each other's types and play a Nash equilibrium of the direct revelation mechanism offered by the principal (see Maskin 1999, and the discussion of the non-verifiability paradigm below). Second, agents may only know their own type, form beliefs on each others' types and play a Bayesian–Nash equilibrium (see D'Aspremont and Gérard-Varet 1979). Third, one may also insist on dominant strategy implementation because it does not depend on the specification of beliefs (see Gibbard 1973; Groves 1973; Green and Laffont 1977). To each implementation concept corresponds a notion of incentive feasibility. Once the set of incentive feasible contracts is defined, one can proceed to optimization. It is a trivial observation that, the more restrictive the implementation concept, the lower is the principal's payoff at the optimum.

In some cases, such as the provision of public goods within a society of privately informed agents or in bargaining models between a buyer and a seller with equal bargaining power, the goal is no longer to design a multilateral contract which would extract the rents of all agents but, instead, to maximize some *ex ante* efficiency criterion under incentive constraints. Groves (1973) showed that dominant strategy mechanisms suffice to implement the first-best decision in a public good context. One caveat is that the budget generally fails to be balanced. D'Aspremont and Gérard-Varet (1979) proposed a Bayesian incentive-compatible mechanism which implements the first-best and still satisfies budget balance. As argued by Laffont and Maskin (1979), such a mechanism may conflict with the agents' participation constraint. In a bargaining environment, Myerson and Satterthwaite (1983) showed in a similar vein

that there exists no Bayesian bargaining mechanism that is efficient, budget-balance and individually rational.

The optimal multilateral contract can be very sensitive to the information structure. In environments where risk-neutral agents have correlated types but know only their own type, the principal can condition one agent's compensation on another's report. By doing so, the principal can fully extract the rent from both agents in a Bayesian-Nash equilibrium. One may view this result as a strong rationale for relative performance evaluation, yardstick competition, benchmarking and internalization of similar activities within the same organization. This puzzling insight of Crémer and McLean (1988) no longer holds when one introduces risk-aversion, *ex post* participation constraints or limited liability constraints. These assumptions reintroduce information rents in the multi-agent organization, and the standard trade-off between efficiency and rent extraction reappears.

When the agents' types are independently distributed, yardstick competition is ineffective and the agents derive information rents. However, the externality that one agent's task may exert on another can shape the distribution of these rents. In competitive environments, such as procurement auctions among sellers, it is no longer the distribution of the agents' marginal costs but the distribution of their virtual marginal costs (see Myerson 1981) which determines who should produce and how much. Because virtual costs may be ranked differently from true costs, inefficiencies arise under asymmetric information. Moreover, competition may help reduce rents by putting each agent under the threat of being excluded from production if he overstates his cost too much. There is then a positive externality among competing agents.

Instead, more cooperative environments, such as public good problems or procurement of complementary inputs by several suppliers, involve negative externalities between agents. Given that each agent has a limited impact on the organization's overall production, the incentives to overstate costs and thereby receive greater transfers are exacerbated. 'Free riding' arises in such organizations (see Mailath and Postlewaite 1990).

When competition between agents or between agents and the supervisors supposed to monitor them would benefit the principal, one must consider the possibility of collusion aimed at securing more rent. Reducing the scope for collusion requires using mechanisms that are less sensitive to information and reducing supervisory discretion. Incentive contracts look more like inflexible bureaucratic rules (see Tirole 1986; Laffont and Martimort 2000). The optimal response to collusion may also entail more delegation to lower levels of the hierarchy, as in Laffont and Martimort (1998) and Faure-Grimaud et al. (2003).

Dynamics

Different extensions of the static framework correspond to different abilities of the contractual partners to commit themselves inter-temporally and/or different ways for the cost parameters to vary over time. Under full commitment, the lessons of the static rent-efficiency trade-off can be easily extended, although the precise features of the optimal contract depend on how types evolve over time (see, for instance, Baron and Besanko 1984, for the case of persistent types). The case of limited commitment is more interesting. Long-term contracts may either be renegotiated (Dewatripont 1989; Hart and Tirole 1988; Laffont and Tirole 1990) or even are not feasible, in which cases the parties resort to spot contracts (Laffont and Tirole 1988). The rent-efficiency trade-off must be adapted to take into account how information is revealed progressively over time. However, the basic idea still holds. As past performances reveal information about the agent's type, the optimal contract trades off *ex post* efficiency gains in contracting against the agent's desire to hide information in the earlier periods of the relationship so as to secure more rent in the later periods.

Applications

Since the mid-1980s, models of optimal contracting under adverse selection have spanned the economic literature. Let us quote only a few major applications. Mirrlees (1971) analysed optimal taxation schemes when the agent's productivity is privately observed. He introduced the Spence–Mirrlees

condition and derived the implementability conditions. He also used optimal control techniques (Pontryagyn Principle) to compute the optimal taxation scheme. (The taxation problem differs from our buyer–seller example because participation in the mechanism is mandatory and the state's budget constraint must be added to the characterization of feasible allocations.)

Mussa and Rosen (1978) studied the problem of a monopolist selling one unit of a good to a continuum of consumers vertically differentiated with respect to their willingness to pay for the quality of this good. This was the first model using adverse selection techniques in a framework without income effect. Maskin and Riley (1984) were interested in characterizing the optimal nonlinear price used by a monopolist in a second-degree price discrimination context.

Baron and Myerson (1982) applied the methodology to the regulation of natural monopolies privately informed about their marginal costs of production. Laffont and Tirole (1986) extended this analysis to allow for cost observability but also introduced moral hazard elements (the possibility for the regulated firm to reduce its costs by undertaking some non-observable effort). They derived cost-reimbursement rules and pricing policies. They showed that menus of linear contracts might implement the optimal contract.

Green and Kahn (1983) and Hart (1983) studied labour market contracts and discussed distortions towards overemployment or underemployment that may arise depending on the contractual environment considered.

Finally, Townsend (1979) and Gale and Hellwig (1985) analysed optimal financial contracts in a framework where the borrower's income is observable only *ex post* and at a cost. Optimal contracts may look like debt in such environments.

Moral Hazard

To return to our buyer–seller example, we now assume that there is only one unit of a good to be traded whose quality q is random and which yields a surplus $S(q)$ to the buyer. The distribution of quality is affected by an effort e undertaken by the agent at a cost $\psi(e)$ (where $\psi' > 0$ and $\psi'' > 0$). The cumulative distribution is $F(q|e)$ (with density $f(q|e)$) on a support $Q\ [\underline{q}, \overline{q}]$ independent of the agent's effort. To simplify, the agent's preferences are separable in money and effort: $U = u(t) - \psi(e)$ where $u(\cdot)$ is increasing and concave ($u' > 0$, $u'' \leq 0$). The agent's outside option is not to produce, which gives him a payoff normalized to zero.

The agent's effort is observable neither by the principal nor by a court of law. This is a *moral hazard* setting. Contracts stipulate the agent's payment as a function of the realized quality assumed to be observable and verifiable (contractible) by a court of law. Therefore, contracts are of the form $\{t(\tilde{q})\}_{\tilde{q} \in Q}$.

If the effort were observable, its value could also be specified by contract. Therefore, the seller can at the same time be forced to exert the first-best level of effort and be fully insured against uncertainty on realized quality with a flat payment independent of his performance:

$$u(t^*) = \psi(e^*).$$

This is no longer the case when the agent's effort is non-verifiable. The first step of the analysis is to describe the set of feasible incentive contracts implementing a given level of effort e.

In a moral hazard setting, incentive constraints write as:

$$\int_{\underline{q}}^{\overline{q}} u(t(q)) f(q|e) dq - \psi(e) \geq \int_{\underline{q}}^{\overline{q}} u(t(q)) f(q|e') dq - \psi(e') \quad \forall (e, e'). \quad (6)$$

The agent's participation constraint is:

$$\int_{\underline{q}}^{\overline{q}} u(t(q)) f(q|e) dq - \psi(e) \geq 0. \quad (7)$$

Risk Neutrality

A first case of interest is when the agent is risk-neutral ($u(t) \equiv t$). The simple 'sell-out' contract, $t(q) = S(q) - C$ where C is a constant, implements

the first-best level of effort e^*. Provided that $\int_{\underline{q}}^{\overline{q}} S(q) f(q|e^*) - \psi(e^*)$, this scheme also extracts all the surplus from the agent who is just indifferent between producing or not.

Intuitively, with such a 'sell-out' contract, the agent's private incentives to exert effort are aligned with the social incentives. This efficient outcome is obtained by, first, having the agent pay a bond worth C for the right to serve the principal, and second, having the principal pay an amount $S(q)$ contingent on the quality realized.

Such a 'sell-out' contract requires that the agent bears the full consequences of a bad performance. It might not be feasible when the agent has limited liability and cannot be punished for bad performances. (For details, see Laffont and Martimort 2002, ch. 4). The conjunction of moral hazard and limited liability allows the agent to derive a limited liability rent. Intuitively, only rewards, not punishments, can be used to provide incentives, and this restriction on instruments is costly for the principal. This rent creates a trade-off between efficiency and rent extraction, as in the adverse selection framework. Effort is distorted below the first-best level.

Risk Aversion

Let us turn to the more complex case of risk aversion. A first concern of the literature has been to 'simplify' the set of incentive constraints (2) by replacing it with a first-order condition:

$$\int_{\underline{q}}^{\overline{q}} u(t(q)) f_e(q|e) dq = \psi'(e). \tag{8}$$

Denoting by λ (resp. μ) the positive multiplier of the incentive (resp. participation) constraint (8) (resp. (7)), the optimal second-best schedule $t^{SB}(q)$ satisfies

$$\frac{1}{u'(t^{SB}(q))} = \mu + \lambda \frac{f_e(q|e)}{f(q|e)}. \tag{9}$$

This condition yields two important insights. First, the contract must simultaneously provide the risk-averse agent with insurance, which requires a fixed payment, and with incentives to exert effort, which requires that payments be linked to performance. There is now a trade-off between *insurance* and *incentives*.

Second, the monotonicity of the agent's compensation with respect to the quality level (a priori a quite intuitive property) is obtained only when the *monotone likelihood ratio property* holds, namely, when $\frac{\partial}{\partial q}\left(\frac{f_e(q|e)}{f(q|e)}\right) > 0$. This property means that higher levels of performance are more informative about the agent's effort.

Finally, the optimal contract must use all signals which are informative about the agent's effort but no uninformative signals. Using them would only let the agent bear more risk without any beneficial impact on incentives. This is the so-called informativeness principle of Holmström (1979).

Extensions

In a model with a finite number of quality and effort levels, Grossman and Hart (1983) offered a careful study of the set of incentive constraints and its consequences for the shape of optimal contracts. There is no general result on the ranking between the first-best and the second-best effort levels in such environments. The discrete version of the first-order approach requires that only nearby constraints matter in the agent's problem. This concavity of the agent's problem is ensured when $F(q|e)$ is itself convex in q. In models with a continuum of effort levels and outcomes, this first-order approach was suggested in Mirrlees (1999), more rigorously justified in Rogerson (1985) and Jewitt (1988) and applied in Holmström (1979) and Shavell (1979).

The moral hazard methodology has been used to justify the optimality of linear incentive schemes in well-structured environments (Holmström and Milgrom 1987); an often found feature of real world contracts. Equipped with this tool, Holmström and Milgrom (1991, 1994) investigated how multiple tasks and jobs should be arranged in an organization.

To avoid the complexity of models with a continuum of effort levels, modellers have found it useful to focus on simplified environments with two levels of effort. This approach was

instrumental in the work on corporate finance of Holmström and Tirole (1997).

Multi-agent Organizations

When applied to multi-agent organizations, the 'informativeness principle' suggests that an agent's compensation should be linked to another's performance if it is informative about his own effort (see Mookherjee 1984). Relative performance evaluation and benchmarking can help eliminate common shocks affecting all agents' performances. Of particular importance in this respect are tournaments which use only the ranking of the agents' performances to determinate their compensations. Tournaments provide agents with insurance against common shocks, which has a positive incentive effect. More generally, the properties of tournaments and how they compare with (a priori suboptimal) linear schemes have been investigated in Nalebuff and Stiglitz (1983) and Green and Stokey (1983).

In more cooperative environments where different agents contribute to a joint project, the fundamental difficulty is how to share the proceeds of production among agents of the team and still provide some incentives. Since each agent enjoys only a fraction of those proceeds but bears the full cost of his effort, he reduces his effort supply. This leads to a free-rider problem within teams, which is analysed in Holmström (1982).

If we remain in cooperative environments but allow now for a principal acting as a budget breaker, this principal may find it worthwhile to reduce the agency cost of implementing a given effort profile by having agents behave cooperatively (Itoh 1993). Even when agents do not cooperate, mutual observability of effort levels can also help to eliminate agency cost, as in Ma (1988). This last argument relies on the logic of non-verifiability models, developed below.

Dynamics

The basic issue investigated by dynamic models of moral hazard is the extent to which repeated relationships alleviate the moral hazard problem. The intuition is that the principal should filter out the agent's effort by looking at the whole history of his performances. This may eliminate any agency problem, at least when parties do not discount too much the future (see Laffont and Martimort 2002, ch. 8, for an example). More generally, the insurance–incentives trade-off may be relaxed when the risk-averse agent's rewards and punishments can be smoothed over the whole relationship, as shown in Spear and Srivastava (1987). A direct consequence of inter-temporal smoothing is that the optimal dynamic contract exhibits *memory*; good (resp. bad) performance today will also affect positively (resp. negatively) future compensations. This insight has been used to formalize a theory of the wage dynamics inside the firm (Harris and Holmström 1982).

Fama (1980) argued that reputation in the labour market exerts enough discipline on managers to alleviate moral hazard even in the absence of explicit contracts. Holmström (1999) built a model of career concerns where the manager's interest in influencing the labour market's beliefs concerning his or her quality provides incentives to exert effort. Career concerns are nevertheless in general not enough to induce first-best effort levels, and some inefficiencies remain.

Non-verifiability

Let us return to the buyer–seller model above. Although we now assume that it is observable by both the principal and the agent, the state of nature θ may still not be verifiable by a court of law, in which case it cannot be part of the contract. This shared knowledge stands in sharp contrast with the asymmetric information structures examined in previous sections.

The first difficulty consists of building a mechanism based only on verifiable variables (namely, the quantities traded and corresponding payments) which implements the first-best quantity $q^*(\theta)$ and transfers $t^*(\theta)$. This problem was addressed by Maskin (1999). He demonstrated that the first-best quantities and transfers can easily be implemented with a direct revelation mechanism $\left\{t\left(\hat{\theta}_a, \hat{\theta}_b\right), q\left(\hat{\theta}_a, \hat{\theta}_b\right)\right\}_{\left(\hat{\theta}_a, \hat{\theta}_b\right) \in \Theta^2}$ where both the buyer and the seller report

simultaneously the state of nature they commonly know. Truth-telling is obviously a Nash equilibrium of this mechanism provided that both traders are severely punished when making different reports, since such cases would be inconsistent with the underlying information structure.

A more subtle issue is how to design a mechanism such that this truthful Nash equilibrium is unique. Maskin (1999) proposed a condition for players' preferences such that this is the case. Moore and Repullo (1988) significantly extended the domain of preferences by hardening the implementation concept, replacing Nash behaviour by subgame-perfection in a sequential moves mechanism (see Laffont and Martimort 2002, ch. 6, for an example, and Moore 1992, for an exhaustive survey of the literature).

The basic thrust of the non-verifiability paradigm is that a court of law can get around non-verifiability by building such revelation mechanisms, at least as long as the non-verifiable state is payoff-relevant. If one sticks to that interpretation, non-verifiability does not present a significant limit on contracting.

A second issue of the literature is the impact of non-verifiability on the incentives of traders to perform specific and non-verifiable investments. Given our previous claim that non-verifiability is generally not a constraint, the model resembles the standard moral hazard model. Providing incentives for investments meets the same difficulties as in the previous section.

Extensions

In practice, revelation mechanisms have been criticized as overly complex, as relying on threats which may either be non-credible or violate limited liability constraints. The so-called incomplete contracts literature has thus focused on cases where such revelation mechanisms are not feasible. In such environments, either no contract at all or only a very rough one can be written *ex ante*. For instance, parties can agree *ex ante* on a simple fixed-price/fixed quantity contract which serves as a threat point for the bargaining which takes place *ex post* when the state of nature is realized (see Edlin and Reichelstein 1996, among others).

Alternatively, this threat point may be determined by the allocation of ownership rights where such a right gives the owner the opportunity to use assets as he prefers in case bargaining fails (see Grossman and Hart 1986; Hart and Moore 1988). The issue is then to derive from those exogenous constraints distortions of investments and optimal organizations which may mitigate those distortions.

The incomplete contracts paradigm is similar to the complete contracts one (adverse selection, moral hazard and non-verifiability) in the sense that it also imposes limits on what a court may verify. It differs from it because it also imposes exogenous restrictions on the set of mechanisms available to the parties. The justification for these restrictions is found either in the bounded rationality of players or the difficulties in describing or foreseeing contingencies, all theoretical issues which remain high on the agenda of economic theorists and are still unsettled. The relevant literature on incomplete contracts is too large to be summarized in this short article. The interested reader may refer to Tirole (1999) for an overview or to the entry for this term in this Dictionary.

See Also

- ▶ Adverse Selection
- ▶ Agency Problems
- ▶ Incomplete Contracts
- ▶ Mechanism Design
- ▶ Mechanism Design (New Developments)
- ▶ Moral Hazard

Acknowledgment I thank D. Gromb and J. Pouyet for helpful comments on an earlier version

Bibliography

Armstrong, M., and J. Rochet. 1999. Multidimensional screening: A user's guide. *European Economic Review* 43: 959–979.

Bagnoli, M., and T. Bergstrom. 2005. Log-concave probability and its applications. *Economic Theory* 26: 445–469.

Baron, D., and D. Besanko. 1984. Regulation and information in a continuing relationship. *Information Economics and Policy* 1: 447–470.

Baron, D., and R. Myerson. 1982. Regulating a monopolist with unknown costs. *Econometrica* 50: 911–930.

Crémer, J., and R. McLean. 1988. Full extraction of surplus in Bayesian and dominant strategy auctions. *Econometrica* 56: 1247–1257.

Crémer, J., F. Khalil, and J. Rochet. 1998. Strategic information gathering before a contract is offered. *Journal of Economic Theory* 81: 163–200.

D'Aspremont, C., and L. Gérard-Varet. 1979. Incentives and incomplete information. *Journal of Public Economics* 11: 25–45.

Dasgupta, P., P. Hammond, and E. Maskin. 1979. The implementation of social choice rules. *Review of Economic Studies* 46: 185–216.

Dewatripont, M. 1989. Renegotiation and revelation information over time: The case of optimal labour contracts. *Quarterly Journal of Economics* 104: 489–520.

Edlin, A., and S. Reichelstein. 1996. Hold-ups, standard breach remedies, and optimal investments. *American Economic Review* 86: 478–501.

Fama, E. 1980. Agency problem and the theory of the firm. *Journal of Political Economy* 88: 288–307.

Faure-Grimaud, A., J.-J. Laffont, and D. Martimort. 2003. Collusion, delegation and supervision with soft information. *Review of Economic Studies* 70: 253–280.

Gale, D., and M. Hellwig. 1985. Incentive-compatible debt contracts: The one-period problem. *Review of Economic Studies* 52: 647–663.

Gibbard, A. 1973. Manipulations of voting schemes: A generalized result. *Econometrica* 41: 587–601.

Green, J., and C. Kahn. 1983. Wage-employment contracts. *Quarterly Journal of Economics* 98: 173–188.

Green, J., and J.-J. Laffont. 1977. Characterization of satisfactory mechanisms for the revelation of preferences for public goods. *Econometrica* 45: 427–438.

Green, J., and N. Stokey. 1983. A comparison of tournaments and contracts. *Journal of Political Economy* 91: 349–364.

Grossman, S., and O. Hart. 1983. An analysis of the principal-agent problem. *Econometrica* 51: 7–45.

Grossman, S., and O. Hart. 1986. The costs and benefits of ownership: A theory of lateral and vertical integration. *Journal of Political Economy* 94: 691–719.

Groves, T. 1973. Incentives in teams. *Econometrica* 41: 617–631.

Guesnerie, R., and J.-J. Laffont. 1984. A complete solution to a class of principal-agent problems with an application to the control of a self-managed firm. *Journal of Public Economics* 25: 329–369.

Harris, M., and B. Holmström. 1982. A theory of wage dynamics. *Review of Economic Studies* 49: 315–333.

Hart, O. 1983. Optimal labour contracts under asymmetric information: An introduction. *Review of Economic Studies* 50: 3–35.

Hart, O., and J. Moore. 1988. Property rights and the nature of the firm. *Journal of Political Economy* 98: 1119–1158.

Hart, O., and J. Tirole. 1988. Contract renegotiation and Coasian dynamics. *Review of Economic Studies* 55: 509–540.

Holmström, B. 1979. Moral hazard and observability. *Bell Journal of Economics* 10: 74–91.

Holmström, B. 1982. Moral hazard in teams. *Bell Journal of Economics* 13: 324–340.

Holmström, B. 1999. Managerial incentive problems: A dynamic perspective. *Review of Economic Studies* 66: 169–182.

Holmström, B., and P. Milgrom. 1987. Aggregation and linearity in the provision of intertemporal incentives. *Econometrica* 55: 303–328.

Holmström, B., and P. Milgrom. 1991. Multi-task principal agent analysis. *Journal of Law, Economics, and Organization* 7: 24–52.

Holmström, B., and P. Milgrom. 1994. The firm as an incentive system. *American Economic Review* 84: 972–991.

Holmström, B., and R. Myerson. 1983. Efficient and durable decision rules with incomplete information. *Econometrica* 51: 1799–1819.

Holmström, B., and J. Tirole. 1997. Financial intermediation, loanable funds, and the real sector. *Quarterly Journal of Economics* 112: 663–691.

Itoh, H. 1993. Coalition incentives and risk-sharing. *Journal of Economic Theory* 60: 416–427.

Jewitt, I. 1988. Justifying the first-order approach to principal-agent problems. *Econometrica* 56: 1177–1190.

Jullien, B. 2000. Participation constraints in adverse-selection models. *Journal of Economic Theory* 93: 1–47.

Laffont, J.-J., and D. Martimort. 1998. Collusion and delegation. *Rand Journal of Economics* 29: 280–305.

Laffont, J.-J., and D. Martimort. 2000. Mechanism design with collusion and correlation. *Econometrica* 68: 309–342.

Laffont, J.-J., and D. Martimort. 2002. *The theory of incentives: The principal-agent model*. Princeton: Princeton University Press.

Laffont, J.-J., and E. Maskin. 1979. A differentiable approach to expected utility maximizing mechanisms. In *Aggregation and revelation of preferences*, ed. J.-J. Laffont. Amsterdam: North-Holland.

Laffont, J.-J., and J. Tirole. 1986. Using cost observation to regulate firms. *Journal of Political Economy* 94: 614–641.

Laffont, J.-J., and J. Tirole. 1988. The dynamics of incentive contracts. *Econometrica* 56: 1153–1175.

Laffont, J.-J., and J. Tirole. 1990. Adverse selection and renegotiation in procurement. *Review of Economic Studies* 57: 597–626.

Laffont, J.-J., and J. Tirole. 1993. *A theory of incentives in procurement and regulation*. Cambridge, MA: MIT Press.

Lewis, T., and D. Sappington. 1989. Countervailing incentives in agency problems. *Journal of Economic Theory* 49: 294–313.
Ma, C. 1988. Unique implementation of incentive contracts with many agents. *Review of Economic Studies* 55: 555–572.
Mailath, G., and A. Postlewaite. 1990. Asymmetric information bargaining problems with many agents. *Review of Economic Studies* 57: 351–638.
Maskin, E. 1999. Nash equilibrium and welfare optimality. *Review of Economic Studies* 66: 23–38.
Maskin, E., and J. Riley. 1984. Monopoly with incomplete information. *Rand Journal of Economics* 15: 171–196.
Maskin, E., and J. Tirole. 1990. The principal-agent relationship with an informed principal. I: Private values. *Econometrica* 58: 379–410.
Maskin, E., and J. Tirole. 1992. The principal-agent relationship with an informed principal. II: Common values. *Econometrica* 60: 1–42.
Mirrlees, J. 1971. An exploration in the theory of optimum income taxation. *Review of Economic Studies* 38: 175–208.
Mirrlees, J. 1999. The theory of moral hazard with unobservable behaviour. Part I. *Review of Economic Studies* 66: 3–22.
Mookherjee, D. 1984. Optimal incentive schemes with many agents. *Review of Economic Studies* 51: 433–446.
Moore, J. 1992. Implementation in environments with complete information. In *Advances in economic theory*, ed. J.-J. Laffont. Cambridge: Cambridge University Press.
Moore, J., and R. Repullo. 1988. Subgame-perfect implementation. *Econometrica* 56: 1191–1120.
Mussa, M., and S. Rosen. 1978. Monopoly and product quality. *Journal of Economic Theory* 18: 301–317.
Myerson, R. 1979. Incentive compatibility and the bargaining problem. *Econometrica* 47: 61–73.
Myerson, R. 1981. Optimal auction design. *Mathematics of Operations Research* 6: 58–63.
Myerson, R., and M. Satterthwaite. 1983. Efficient mechanisms for bilateral trading. *Journal of Economic Theory* 28: 61–73.
Nalebuff, B., and J. Stiglitz. 1983. Prizes and incentives: Towards a general theory of compensation. *Bell Journal of Economics* 14: 21–43.
Rogerson, W. 1985. The first-order approach to principal-agent problems. *Econometrica* 53: 1357–1368.
Shavell, S. 1979. On moral hazard and insurance. *Quarterly Journal of Economics* 93: 541–562.
Smith, A. 1776. *The wealth of nations*, 1991. New York: Prometheus Books.
Spear, S., and S. Srivastava. 1987. On repeated moral hazard with discounting. *Review of Economic Studies* 54: 599–617.
Spence, M. 1973. Job market signaling. *Quarterly Journal of Economics* 87: 355–374.
Spence, M. 1974. *Market signalling: Informational transfer in hiring and related processes*. Cambridge, MA: Harvard University Press.
Tirole, J. 1986. Hierarchies and bureaucracies: On the role of collusion in organizations. *Journal of Law, Economics and Organization* 2: 181–214.
Tirole, J. 1999. Incomplete contracts: Where do we stand? *Econometrica* 67: 741–782.
Townsend, R. 1979. Optimal contracts and competitive markets with costly state verification. *Journal of Economic Theory* 21: 417–425.

Contracting in Firms

Canice Prendergast

Abstract

This article provides an overview of recent advances in theoretical and empirical work on incentive contracting in firms. The specific focus is on a variety of reasons why the prediction of the early literature on contracting –suggesting a strong relationship between performance and pay – has not been borne out.

Keywords

Agency theory; Contracting in firms; Externalities; Free rider problem; Incentive contracts; Input monitoring; Measurement error; Multitasking; Performance-related pay; Principal and agent; Private information; Risk

JEL Classifications

J410

In many realms of economic life, the actions of individuals affect the welfare of others. Nowhere is this more relevant than in firms, where employees act on behalf of owners or shareholders to provide services for customers and clients. This separation of the interests of employees from those whose actions they benefit has generated a large literature on incentive contracting, where the overarching objective is the alignment of such interests. The early literature on agency theory, described in the first edition

of this volume by Lazear (1987), conceptually mimics that on externalities – the other area of economics that deals with welfare consequences of actions on others – by showing a variety of ways in which the compensation of agents can be constructed to internalize the effects on one's actions on others. There are two ways of doing this. First, one could simply tell employees what to do and to penalize them if they fail to do so. In the literature, this is referred to as input monitoring. While this can sometimes help, it is often hard to monitor either what workers do, or the intensity with which they do so – a salesman on the road would be a good example. Similarly, while overseers can sometimes identify what it is that agents are doing, they may not know what they *should* be doing – a board of directors monitoring a CEO would be apposite here. Accordingly, the second solution to misaligned incentives is to design compensation plans such that the agent's pay depends on her contribution – 'output' – so that the concerns of other parties are internalized.

A simple model can illustrate this point, and is useful to describe other complications that can arise. The agent is assumed to take some action ('effort') $e \geq 0$, which is unobserved by the principal. She is averse to exerting effort. Consider a simple parameterization of the agent's utility function, where the agent cares about wages w and effort; assume that the agent has exponential utility $V = -\exp[-r(w - C(e))]$, where w is the worker's wage, $r \geq 0$ is the constant rate of absolute risk aversion, the worker's cost of supplying effort is $C(e) = \frac{ce^2}{2}$, and her reservation utility is U^*. To focus attention on the role of output contacting, assume that the principal cannot observe effort e (so monitoring of inputs is not possible), but instead only observes a signal on effort $y = e + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2)$, with σ^2 representing measurement error. Assume also that the principal chooses to reward the agent in a linear fashion on output – a piece rate: $w = \beta_0 + \beta_{yy}$ (There is a large literature on the optimal shape of compensation contracts – see Prendergast 1999; Gibbons 1996, for an overview). Then there is a simple solution to attaining efficient effort: choose the contact to internalize the benefit to others by setting $\beta_y = 1$. In words, efficiency arises when the agent is residual claimant on the benefit of others.

This solution, providing a simple prescription for how compensation contacts should be designed, is both simple and intuitive. And empirically false. There are, of course, some occupations where one can find evidence of such 'high-powered incentives', where agents are essentially residual claimants on output. Indeed, the literature on agency theory is replete with references to such occupations – taxi cab drivers, franchisees, sharecropper farmers and the self-employed. Yet these are exceptions; instead, 'low-powered incentives' in firms are more the norm (see Prendergast 1999, for details). Consequently, one of the quandaries of the literature has become why so few workers seem to have contracts where their pay is strongly linked to their performance, and much of the subsequent literature to that outlined in the first edition of the *New Palgrave* has identified relevant constraints on incentive contracting.

The earliest candidate to explain why high-powered incentives are rare is that high-powered contracts impose *risk* on workers (Holmstrom 1979). Consider the contract that induces efficient effort above: $\beta_y = 1$. The objective of the firm is to maximize profits subject to the worker's willingness to take the position. This implies that the fixed component, β_0, is changed to guarantee that agents earn their reservation utility, so the principal's objective becomes a surplus maximization exercise. When the worker is risk neutral, the fixed component is reduced sufficiently such that the total compensation cost is $U* + \frac{c}{2}$. In words, the only cost that the employer incurs in addition to U^* is the effort cost. This is not true when the worker is risk averse. In the context of the preferences V above, compensation costs increase when incentive contracts are used for two reasons – the cost of increased effort as above, but also a risk cost imposed on workers. Both costs are increasing in β_y. With exponential preferences and linear contracts, this trade-off results in the optimal contract being $\beta_y^* = \frac{1}{1+rc\sigma_y^2}$. This approach to studying incentive contracting has become knows as the 'trade-off of risk and

incentives', where firms trade off the benefits of great effort with higher compensation costs induced by a risk premium, such that the chosen level of effort falls below the level that internalizes benefits to others. Only in the case where there is either no measurement error ($\sigma_y^2 = 0$) or risk neutrality ($r = 0$) does efficient effort arise.

At its most general, this costliness of exposing a worker to large degrees of risk (or its analogue, liquidity constraints) surely explains some part of the absence of high-powered incentives. In much the same way as financial assets with higher undiversifiable risk require higher expected returns, so also are risky jobs likely to demand higher compensation. Despite this, the empirical literature on how compensation contracts trade off such risk issues against higher effort has shown little evidence in its favour. There are two principal empirical implications of the theory. First, riskier environments should have lower incentives $-\beta_y$* declines with σ_y^2. There have been many studies of the relationship between risk and the strength of incentives in a variety of occupations. If anything, this literature suggests that the relationship between risk and the provision of incentives is positive rather than the negative relationship posited by this theory. See Prendergast (2002) for details and an explanation as to why this may be. Second, the trade-off of risk and incentives implies that compensation should not depend on measures that workers cannot control. Again, this has found little support in the data. For example, Bertrand and Mullainathan (2001), have examined executive contracts in the United States, and found little evidence that contracts reward executives any less for measures that they cannot control (say, where an oil company's profits change simply because the price of crude changed) than for those that they can (such as a merger). More evidence on this failure to filter out uncontrollable factors concerns the infrequency of relative performance evaluation. Consider two sales-force workers (or executives) who carry out a similar job. If demand for the products that they sell varies for common reasons beyond their control, an efficient way of limiting risk exposure is to (at least partially) reward the workers on how well they do relative to each other. Yet empirically there is relatively little evidence of such benchmarking (for example, see Janakiraman et al. 1992).

A second limitation on incentive contracting arises when measures do not reflect the objectives of the principal. Workers often carry out a host of activities in their jobs, yet measures of performance may not reflect all these aspects. A good example of this would be measuring the performance of a teacher. While measures may be available on some component of what they do – such as test scores for a teacher – many important aspects may remain unmeasured. When contracts are designed on the subset of things that can be measured, there is a danger that they ignore the unmeasured aspects. For instance, there is evidence of teachers 'teaching for the test' or cheating to achieve higher test scores (Jacob and Levitt 2003). This phenomenon has become known as *multitasking* (Holmstrom and Milgrom 1992), which becomes potentially important when there is no single measure that reflects the contribution of an agent. Accordingly, it is not surprising that a consistent empirical finding is that jobs which are described by firms as complex tend not to offer significant incentive pay (see Prendergast 1999, for details).

Another limitation on the ability of firms to provide incentives to workers comes from team production. Measures of performance for most workers reflect not only what they do but also the contributions of others. In itself, this does not change the calculus above in any conceptual sense, other than that the measurement error now includes the actions of others. As an example, assume that two agents (1 and 2) work on a team and that output measures the true contributions of both plus an error term $y = e_1 + e_2 + \varepsilon$. Efficient effort arises as before by setting $\beta_y = 1$ for each worker. However, there is now a potential problem of budget breaking, where marginal payments exceed marginal output. In this example, when total output rises by one dollar, compensation costs increase by two dollars. In many firms – for instance, partnerships – such budget breaking is not possible. If instead the principal

can pay out no more than one dollar for every dollar extra on output, this naturally places an upper bound of $\beta_y = \frac{1}{2}$ on average for the agents. Hence, budget balancing places a natural limitation on firm incentives. This also leads to a free rider problem in teams, where maximum incentive compensation in an N member team mechanically declines as N increases (This is known as the '$\frac{1}{N}$problem'). There is also considerable empirical evidence (such as Gaynor and Pauly 1990) on such free riding – mostly from legal and medical partnerships – illustrating how various measures of performance disimprove as the size of the team being rewarded increases.

Many measures of output are not denominated in dollar terms, but instead come in the form of evaluations by others. For instance, it would be difficult to measure the contribution of a social worker or a customer service representative without using feedback from supervisors or clients. Another limitation on contracts arises when such subjective measures can be corrupted by evaluators with vested interests. Two particular sources of such vested interests have been considered in the literature. First, information on performance often originates with clients as they are the only ones with first-hand experience of the agent's efforts. For instance, compensation for many customer service representatives depends on client evaluations. When clients have relatively similar preferences to the principal – such as that the agent should be courteous and efficient – contracts based on evaluations can mimic the objective contracts above. Yet in other instances, the vested interest of clients can render incentives difficult to implement. A good example of this arises in occupations such as police or immigration control, whose objective is not necessarily to make their clients happy. Making pay depend on evaluations in these instances can be harmful as it gives agents incentives to keep clients happy when they should not, such as a police officer not arresting a suspect. In these cases, incentive contracting on evaluations typically needs to be curtailed to avoid such incentives (see Prendergast 2003, for details).

The second example of vested interests with subjective evaluations is where the principal has an incentive not to implement the *(ex ante)* efficient contract by reneging on a promised payment to save costs. Thus, even though an agent exerts effort and performs well, the supervisor claims otherwise to keep costs down. This can arise either by outright lying or perhaps by manipulating whatever measures are available. A relevant example here is the movie industry, where actors are sometimes paid on the 'net profits' of a film. As a result, there have been numerous court cases regarding firms using creative transfer pricing arrangements to reduce profits for very successful movies. See Cheatham et al. (1996) for more details on this. Such incentives to renege are likely even worse when there are no objective measures of performance. Because the desire to renege is greater when discretionary incentive payments are higher, it follows that the only credible contracts often involve few incentives (Clive Bull 1987, considers a role for repeated interaction between the principal and agent as a means of reducing incentives to renege. While repeating the relationship can result in sufficient incentives for complete honesty by the principal, it remains the case that, if the relationship's value is not sufficiently great, incentives must be muted to reduce incentives for cheating).

It is incorrect to assume that the ability to manipulate measures of performance always mute incentives – sometimes it can result in incentive pay being inefficiently high. Consider again two occupations where agents are typically residual claimants – taxi drivers and sharecroppers. At first blush, it would seem odd that they have such extreme incentives. Aren't these as likely candidates for trading off risk against incentives as any? However, one characteristic of each of these occupations is that they have opportunities for hiding output, either by taking fares without using the metre (in the case of cab drivers) or selling crops privately (in the case of farmers). In both cases, the only outcome that makes this incentive irrelevant is to render them residual claimants, even if risk considerations would suggest otherwise.

Another issue which can constrain efficient incentives, yet which has received almost no attention in the empirical literature, is where

agents hold *private information.* Take a specific instance – real estate agents. In Chicago, real estate contracts take a simple form – agents make three per cent of the sales price of the house. Assume that my home is worth $500. This linear contract not only offers only three per cent on the relevant margin for improving the selling price of the house, but predominantly rewards the agent for selling the house for say $450. Yet anyone could sell the house for $450 and it seems highly inefficient to reward in this way. So why not renegotiate to something better? An example of such an improvement (subject to risk issues) would be to offer nothing on the first $450, but to pay a piece rate of 30 per cent on anything over $450. In this way, the agent has more incentives on margin, yet breaks even relative to the original contract if the house sells for its original price.

One reason why such renegotiation does not arise is that the agent may privately know the true value of the home, while the owner believes it to be worth $500 on average. Consider a homeowner who offers the new contract above to the agent. It is clear that the agent rejects the new contract if it is truly worth less than expected, and accepts it if worth more. But this implies that the agent earns *information rents* on average. As a result, on average the homeowner loses money from the renegotiation unless effort increases enough. This option available to the agent limits the ability of contracts to attain efficiency. Instead, in the usual monopoly fashion, the homeowner would offer a contract to trade off the efficiency gains of increased effort with infra-marginal losses of the type described above, resulting in lower-powered incentives (There is a large mechanism design literature on this topic that has largely been ignored by the empirical literature on incentives; see Laffont and Tirole 1986, for example. This is surely partly because of the empirical conundrum as to why mechanisms are so rare in reality).

Much of the recent literature has been focused on how incentive contracts can cause adverse behavioural responses. Another possible mechanism for such responses is where *intrinsic motivation* can be crowded out by the use of incentive contracts. The premise of this literature has been that in many occupations agents enjoy carrying out the activity or care about the outcomes of their actions. As a result, they will exert effort beyond that which they can get away with even in the absence of incentive contracts. This, in itself, is not enough to limit incentive contracting. However, there is some psychological evidence that agents enjoy their jobs less when incentive contracting is used. In effect, they feel that they are only doing it 'for the money' and hence lose interest. A commonly cited example is the willingness of people to donate blood, where the warm feeling from donating declines when payments are made. In some instances, this can imply that incentive contracting can *reduce* effort if these crowding out effects are strong enough. As a result, it can be optimal to provide no incentives even when effort is one-dimensional. This area of research, whose empirical testing has largely been restricted to the laboratory, is still in its early stages and is likely to see much refining over the coming years. See Frey and Jegen (2001) for a survey.

Another likely fruitful area of future research concerns non-monetary ways of motivating workers. This literature largely began as an exercise in how workers could be motivated to internalize the benefits of others, yet has almost exclusively become an exercise in how to motivate through monetary contracting. Yet it is clear that there are a myriad of means of motivating workers – sense of achievement, 'doing good', status, and so on – that firms tap into. How such mechanisms operate, and the way in which they interact with monetary contracts, remain an unstudied topic of research, though see Besley and Ghatak (2005), for some theoretical work on this issue.

It is worthwhile to note a caveat before concluding. The discussion above concerns the absence of *observed* incentive contracts. Yet workers often have unobserved carrots and sticks that can motivate them. For instance, many workers exert effort in the hope of attaining a promotion (Lazear and Rosen 1981), or a better job offer (Holmstrom 1999). Many of these mechanisms for inducing desired behaviour are dynamic, where good performance today results

in a greater likelihood of promotion, or better job offers in future. Such incentives are clearly important for workers. However, it remains the case that explicit incentive payments remain limited even in those cases where the above types of career concern are negligible (For example, it is well known that promotion prospects become very limited for workers who remain in a job grade for a long period. Yet explicit incentives are no more common for those workers than for any other). The interaction of unobserved (typically career) incentives with the more explicit set of piece rates and bonuses that have been considered above is surely of first-order importance to firms, though it remains surprisingly unexplored in the literature (see Baker et al. 1994, for an exception).

To conclude, perhaps the central foundation of modern economics is the idea that appropriate prices guide behaviour in efficient ways. Despite this, one of the defining characteristics of the employment relationship in many firms is the absence of the kind of explicit prices whereby wages depend in a clear way on observed outcomes. The early incarnations of agency theory were concerned with designing prices in a way that could serve to fully internalize the effects of agents' actions on the welfare of their employers. Yet this initial optimism has now been tempered with a somewhat more nuanced view that shows trade-offs that will ultimately help in defining more precisely the nature of labour market relationships.

Bibliography

Baker, G., R. Gibbons, and K.J. Murphy. 1994. Subjective performance measures in optimal incentive contracts. *Quarterly Journal of Economics* 109: 1125–1156.

Bertrand, M., and S. Mullainathan. 2001. Are CEOs rewarded for luck? The ones without principals are. *Quarterly Journal of Economics* 116: 901–932.

Besley, T., and M. Ghatak. 2005. Competition and incentives with motivated agents. *American Economic Review* 95: 616–636.

Bull, C. 1987. The existence of self-enforcing wage contracts. *Quarterly Journal of Economics* 102: 147–159.

Cheatham, C., D. Davis, and L. Cheatham. 1996. Hollywood profits: Gone with the wind? *CPA Journal* 12: 32–34.

Frey, B., and R. Jegen. 2001. Motivation crowding theory: A survey of empirical evidence. *Journal of Economic Surveys* 15: 589–611.

Gaynor, M., and M. Pauly. 1990. Compensation and productive efficiency in partnerships. Evidence from medical group practice. *Journal of Political Economy* 98: 544–574.

Gibbons, R. 1996. Incentives and careers in organizations. In *Advances in economics and econometrics: Theory and applications*, ed. D. Kreps and K. Wallis. Cambridge: Cambridge University Press.

Holmstrom, B. 1979. Moral hazard and observability. *Bell Journal of Economics* 10: 74–91.

Holmstrom, B. 1999. Managerial incentive problems: A dynamic perspective. *Review of Economic Studies* 66: 169–182.

Holmstrom, B., and P. Milgrom. 1992. Multi-task principal agent analyses: Linear contracts, asset ownership and job design. *Journal of Law, Economics, and Organization* 7: 24–52.

Jacob, B.A., and S.D. Levitt. 2003. Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *Quarterly Journal of Economics* 118: 843–877.

Janakiraman, S.N., R.A. Lambert, and D.F. Larker. 1992. An empirical investigation of the relative performance evaluation hypothesis. *Journal of Accounting Research* 30: 53–69.

Laffont, J.-J., and J. Tirole. 1986. Using cost observation to regulate firms. *Journal of Political Economy* 94: 614–641.

Lazear, E.P. 1987. Incentive contracts. In *The New Palgrave: A Dictionary of Economics*, ed. J. Eatwell, M. Milgate, and P. Newman, vol. 2. London: Macmillan.

Lazear, E., and S. Rosen. 1981. Rank order tournaments as optimal labor contracts. *Journal of Political Economy* 89: 841–864.

Prendergast, C. 1999. The provision of incentives in firms. *Journal of Economic Literature* 37: 7–63.

Prendergast, C. 2002. The tenuous trade-off between risk and incentives? *Journal of Political Economy* 110: 1071–1102.

Prendergast, C. 2003. The limits of bureaucratic efficiency. *Journal of Political Economy* 111: 929–959.

Contradiction

Michael Dummett

The fundamental form of a contradiction is a pair of propositions, 'A' and 'Not A', one the negation of the other. If such an explicit contradiction is

part of a body of propositions asserted by some individual or group at a given time, it follows that not all those propositions can be true: by thus impairing the reliability of the proponent, the occurrence of the contradiction throws doubt upon the truth of all the other propositions asserted.

Far more frequent than an overt or explicit contradiction is a hidden contradiction. A hidden contradiction is contained in a body of propositions when, by logically valid deductive reasoning, an explicit contradiction can be derived from them. Two quite different cases arise, according to the status of the original propositions from which the contradiction was derived. The first is that in which, as before, those propositions were all asserted by an individual or body of people. In this case, the hidden contradiction is as fatal to the joint correctness of those assertions as is the explicit one, although, of course, considerable work may have had to be expended in bringing it to light.

The second case is that in which at least some of the original propositions were not asserted, but merely advanced as suppositions to be considered. The formal conclusion now remains exactly the same: given that the contradiction was validly derived, not all the original propositions can be true. The effect, of course, is very different. Suppose that the original propositions – the premisses from which the contradictory pair 'A' and 'Not A' have been derived – were four in number: 'B_1', 'B_2', 'C_1', and 'C_2'; and suppose that 'B_1' and 'B_2' were asserted outright, but that 'C_1' and 'C_2' were merely advanced for consideration. The contradiction shows that not all four can be true: without having to withdraw anything that he asserted, the proponent, still maintaining 'B_1' and 'B_2', is now in a position to assert 'If C_1, then not C_2' (or its equivalent, 'If C_2, then not C_1').

For this reason, derivation of a contradiction can be employed, not merely as a means of refuting the assertions of another, but as a method of demonstrating negative propositions: this is the celebrated mode of argument *reductio ad absurdum*. In conjunction with premisses, 'B_1' and 'B2', say asserted outright, a proposition 'C' is presented as a hypothesis: not as a conjecture or supposition to be seriously entertained, but with an eye to proving its negation 'Not C'. From the premisses 'B_1' and 'B_2' and the hypothesis 'C', two contradictory propositions 'A' and 'Not A' are then derived: on the basis of the two premisses, the hypothesis now being dropped, its negation 'Not C' can now be definitely asserted. Two special cases fall under this general description. One is that under which 'A' is actually one of the two premisses, say 'B_1'. If from premisses 'B2' and the hypothesis 'C', the conclusion 'Not B', can be derived, one may assert 'Not C' on the strength of the two premisses. The other special case is that in which 'A' is the hypothesis 'C', itself. If from premisses 'B_1' and 'B_2' and the hypothesis 'C', the negation 'Not C' of the hypothesis can be derived, 'Not C' may be asserted outright on the strength of the two premisses.

As is well known, *reductio ad absurdum* arguments are frequent in mathematics. One of the simplest, as well as historically most important, is the proof that 2 has no rational square root. We may take as the premisses: (B_1) any rational number may be represented as a fraction whose numerator and denominator have no common factor; and (B_2) the square of an odd number is odd. As the hypothesis we may take: (C) 2 has a rational square root. From (B_1) and (C) it follows that, for some integers m and n, m and n have no common factor and $(m/n)_2 = 2$; hence $m^2 = 2n^2$. Hence, by the definition of 'even', m_2 is even, and from (B_2) it can be inferred that m is not odd, and hence is even. Thus $m = 2k$ for some k, and we have: $m_2 = 4k^2 = 2n^2$, and so $2k^2 = n^2$. Applying (B_2) once more, we infer that n is also even. Since m and n are both even, they have 2 as a common factor, which contradicts the earlier stipulation that they have no common factor. By arriving at this contradiction, we have achieved a demonstration that 2 has no rational square root, that is, a proof of the negation of our hypothesis (C).

The derivation of a hidden contradiction from premisses all definitely asserted may also be illustrated from mathematics. The most celebrated examples are the set-theoretic paradoxes which provoked the first 'crisis in the foundations' of mathematics in the early years of this century. Some of these, such as Burali-Forti's paradox and Cantor's paradox, demand some technical background; but Russell's paradox has the advantage of being statable in very simple terms. The propositions leading to this contradiction appear at first sight entirely harmless. They are: (1) to every property there corresponds a class, the class of things having that property; (2) the class of things having a given property F contains as members all and only those things that have the property F. Russell's contradiction arises from considering the property G of being a class that does not contain itself as a member. By assumption (1), there exists a class, which we may call W, of things having the property G. We now ask whether W contains itself as a member. It may then be argued by *reductio ad absurdum* that it does not: we have first to suppose, as a hypothesis, that it does. By assumption (2), it therefore has property G. But properly G is the properly of being a class that does not contain itself: it follows, contrary to the hypothesis, that W does not contain itself. We may thus discard the hypothesis, and assert outright that W does not contain itself. Applying assumption (2) once more, we may then conclude that W lacks property G. It therefore is either not a class or does contain itself as a member; since it is by definition a class, it contains itself as a member. This now contradicts our earlier conclusion that it does not contain itself as a member: the two intuitively reasonable assumptions (1) and (2) have led to a contradiction.

Russell's contradiction is so simple to state and so easy to derive that those who are unaware of the importance in mathematics of the notion of a class or set are liable to think it no more than a trivial verbal puzzle. It is very far from trivial. It overthrew the life's work of the great logician Gottlob Frege, namely to provide unquestionably firm foundations for the theories of natural numbers and of real and complex numbers, just when he believed that he had accomplished it; and it cast the foundations of mathematics into confusion for some years. By showing that assumptions (1) and (2) cannot both be consistently made, it challenged mathematicians and logicians to find weaker assumptions that would be consistent: for which properties F do assumptions (1) and (2) hold good? This question proved exceedingly difficult to answer: hence the 'crisis in the foundations'.

Why is the discovery of a hidden contradiction so devastating? One answer is that, if a pair of contradictory propositions are both consequences of a given set of assumptions, then *every* proposition is a consequence of them. This holds good in virtue of the logical law, known to the medievals as *ex falso quodlibet*, that, from a pair of premisses 'A' and 'Not A', any proposition 'B' may be inferred. This law appears at first glance both useless and implausible; a natural first reaction is therefore to suggest repudiating it. That, however, is not easily done, since it is a consequence of two other, seemingly inescapable, laws concerning the logical constant 'or'. The first is the law now usually known as 'or'-introduction or disjunction-introduction, that, for any propositions 'A' and 'B', 'A or B' follows from 'A'. The second is that known in the traditional logic as *modus tollendo ponens*, that 'B' follows from the premisses 'A or B' and 'Not A'. It is quite obvious that, from the premisses 'A' and 'Not A' of the *ex falso quodlibet*, its conclusion 'B' can be reached by first applying 'or'-introduction and then *modus tollendo ponens*: but it is very hard to see how we could possibly reject either of the two latter laws, for they appear absolutely constitutive of the meaning of the connective 'or'. The best possible way to establish the truth of a disjunction – a statement of the form 'A or B' – is to establish the truth of one or other of the disjuncts 'A' and 'B': if the truth of 'A' does not guarantee the truth of 'A or B', what does 'or' mean? Conversely, the truth of a disjunction appears to *demand* that at least one of the disjuncts be true: so, if 'A or B' is

true and 'A' is not true, 'B' must be true. If 'Not A' is true, 'A' cannot be true: so the validity of *modus tollendo ponens* appears likewise unquestionable.

Faced with the difficulty of rendering contradictions harmless by modifying our logic, some have proposed extruding the sign of negation altogether from the language. What use are merely negative propositions, they enquire: when we are tempted to assert one, we usually have to hand a more informative affirmative statement to take its place, and, when we do not, it will be a good thing to search for one. But the proposed remedy, drastic as it sounds, does not work: for, in place of the derivation of a contradiction, it is usually possible to exhibit a more direct method of deriving any proposition whatever. This may be illustrated by converting the Russell paradox into this form. Take any random proposition, say 'The Earth is flat': and consider the property H of being a class such that, if it contains itself as a member, the Earth is flat. By assumption (1), there exists a class, say Y, of things having the property H: let us ask, as before, whether Y contains itself as a member. Suppose, as a hypothesis, that it does. Then, by assumption (2), it possesses property H: that is, it is a class such that, if it contains itself as a member, the Earth is flat. Now, by hypothesis, it *does* contain itself as a member; hence, on this hypothesis, the Earth must be flat. We have now shown, on the hypothesis that Y contains itself as a member, that the Earth is flat: we may therefore assert outright, independently of the hypothesis, that, *if* Y contains itself as a member, the Earth is flat. From this we see that, since Y is by definition a class, it has property H. Hence, by assumption (2), Y contains it as a member: that is, Y really does contain itself as a member (and not just by hypothesis). We have now shown two things: first, that if Y contains itself as a member, the Earth is flat; and, secondly, that Y *does* contain itself as a member. It follows inescapably that the Earth is flat.

This version of the paradox lies in wait for any who would seek to escape it by tampering with the logical operation of negation. The idea, that some have had, that escape lies in rejecting the principle of bivalence (that every proposition is either true or false) or the closely related law of excluded middle (that, for any proposition 'A', 'A or not A' is a logical truth), does not even avoid the original paradox. It is tempting to open the argument by which the contradiction is derived with the declaration, 'Either W contains itself as a member or it does not': but it is quite unnecessary, and the foregoing statement of the argument neither began in this way, nor invoked the law of excluded middle at any other point. It turned principally on an application of *reductio ad absurdum*, in order to infer 'W does not contain itself as a member' from the fact that a contradiction followed from the hypothesis 'W contains itself as a member'. There are, indeed, logical systems in which the law of excluded middle does not hold. But, in the principal logical system of this kind, that known as intuitionistic logic, *reductio ad absurdum*, in the above form, is perfectly valid: indeed, all arguments and forms of argument hitherto considered are intuitionistically valid. It is true, indeed, that there is another, frequently used, version of *reductio ad absurdum* that is *not* intuitionistically valid. As *reductio ad absurdum* was characterized above, it always leads to a *negative* conclusion: the derivation of a contradiction from a hypothesis shows that hypothesis is *not* true. In the intuitionistically invalid version, a proposition is deduced to be true from the fact that the hypothesis that its negation is true leads to a contradiction. Certain propositions, say 'B_1' and 'B_2', are asserted as premisses, and 'Not C' is assumed as hypothesis. A contradiction 'A' and 'Not A' is derived, and, on the strength of this, 'C' is asserted as following from the premisses 'B_1' and 'B_2' alone. *This* form of argument does depend, for its validity, on the principle of bivalence. If we equate the falsity of 'C' with the truth of 'Not C', then the contradiction shows that 'C' cannot be false: to infer that it is actually true requires the presupposition that it is either true or false. Intuitionistic logic sanctions our treating large classes of propositions as satisfying bivalence, so that many

applications of this extended form a *reductio ad absurdum* will be legitimate: but it objects to considering it generally valid, since not all propositions can be regarded as necessarily being either true or false.

The devastating effect of the appearance of a hidden contradiction in a set of propositions asserted as true, or hitherto accepted as true, is not, of course, due to the fear that anyone will use the *ex falso quodlibet* law to deduce arbitrary conclusions: the utility of the *ex falso quodlibet* turns on its use in subordinate deductions (deductions under a hypothesis subsequently to be abandoned). The point is, rather, that we no longer have *any* reason to believe a proposition on the strength of its being a logical consequence of the given set, since every proposition is such a consequence: until we have discovered what gave rise to the contradiction, and have corrected it, no conclusion we derived from that set has any claim to be believed. The great philosopher Ludwig Wittgenstein scoffed at mathematicians for their 'superstitious awe and dread in face of a contradiction', and asked why they did not simply go round it: but, until it has been resolved, no one can be sure that the path he has taken does go round it. Analysis may show that the use of one particular notion is responsible for the contradiction, and we may then trust those conclusions, previously drawn, the argument to which in no way involved that notion: but, once the contradiction has appeared, only strong measures will supply us with any rational ground for believing any proposition that does involve it, directly or in the argument it is based on. It is not enough merely to find a way to weaken those of our former assumptions that involved the suspect notion so as to block the derivation of the contradiction. That was precisely what Frege did when Russell showed that his contradiction could be derived in Frege's logical theory: but it proved that another contradiction, similar though more complex, was lurking in the modified theory. We have, therefore, in such a case, not merely to weaken our assumptions, but to supply some argument that at least makes it plausible that the weakened assumptions are now consistent. When no hidden contradiction has been revealed, intuitive acceptability may be sufficient basis on which to treat our assumptions as true, even though it does not amount to proof. Once a contradiction has appeared, however, intuition is no longer to be trusted. The contradiction is evidence that the theory is diseased. It is no more enough to root out that particular contradiction than it is to eliminate a particular patch of dry rot: a thorough decontamination of the whole is called for.

The reason why the appearance of a contradiction is so dire a symptom lies not so much in any laws of inference as in the semantic notions of truth and falsity. A proposition and its negation cannot both be true, whereas it is inherent in the definition of 'follows (logically) from' that what follows from a set of true propositions must itself be true: hence, if 'A' and 'Not A' both follow from a set of assumptions, not every assumption in that set can be true. Appeal to semantic notions is in fact needed to explain the concept of the negation of a proposition. Logicians are accustomed to employ a negation operator, with the sense 'It is not the case that . . .', which acts on whole propositions (sentences) to form their negations; and that has hitherto been tacitly done in this article. But natural language hardly possesses such an operator. Apart from the clumsy expedient of sticking the phrase 'It is not the case that . . . in front of a sentence, natural language has no uniform way of forming negations. Admittedly, the negations of many sentences can be found by negating the main verb (in English, replacing 'is' by 'is not', 'resembles' by 'does not resemble', etc.), but this does not work for all. The negation of 'Someone is snoring' is not 'Someone is not snoring', but 'No-one is snoring', the negation of 'You must attend' is not 'You must not attend', but 'You need not attend'; the negation of 'Whenever I look up, I see you yawning' is not 'Whenever I look up, I do not see you yawning', but 'I do not see you yawning whenever I look up'; and the negation of 'If taxes are cut, the government will be re-elected' is not 'If taxes are cut, the government will not be re-elected', but 'The government

will not necessarily be re-elected if taxes are cut'. There is no uniform syntactic rule for determining the negation of any given proposition: to identify the negation of a proposition 'A', we have to advert to its characterization as that proposition which (in any given context) is true just in case 'A' is false.

This means that the notion of falsity is essential for an explanation of 'not' as used to qualify whole sentences: it does not mean that the notion is more fundamental than the uses of 'not' that occur in natural language. On the contrary, to identify the negations of most propositions, it is necessary to understand the word 'not' as it occurs in a variety of positions in a sentence, as illustrated in the above examples. It is therefore in no way inconsistent with the general characterization of the negation of a proposition in terms of truth and falsity to explain a proposition as being false just in case it is not true. It is, in fact, a mistake to think of truth and falsity as equally fundamental notions: the application to a proposition of only one of them is sufficient to determine its content. If this were not so, there could be one pair of propositions 'A' and 'B' differing only in that 'B' was neither true nor false in certain of the cases in which 'A' was false, and another pair 'C' and 'D' differing only in that 'D' was neither true nor false in certain of the cases in which 'C' was true, and these differences would entail that the two propositions in each pair diverged in content. But that is impossible, if 'content' is understood as here intended, namely as what is conveyed by the assertion of the proposition on its own. What was not laid down, in the foregoing stipulation, was whether, in asserting a proposition, a speaker is to be understood as excluding or as allowing for the case in which that proposition is neither true nor false. If he is to be understood as excluding it, what he conveys by his assertion is that the case in which the proposition is true obtains: the content of 'A' and of 'B' will therefore coincide. If, on the other hand, the speaker is to be understood as allowing for the case in which the proposition is neither true nor false, what he conveys by asserting it is that one of the cases in which it is not false obtains: if so, the content of 'C' and 'D' will coincide. To fix the content of a proposition, in the sense explained, we need to know, of any state of affairs specified in sufficient detail, whether an assertion of the proposition would rule it out or allow for it: and this does not provide for the introduction of two notions, those of truth and falsity, whose application is partially independent of one another.

The principle of bivalence rests on the conception that, to grasp the content of a proposition, one must know just what circumstances must obtain for it to be true, independently of whether or not these are circumstances which we are capable of recognizing as obtaining when they do. Given this conception, we may understand 'Not A' as being true in all circumstances save those in which 'A' is true: and so 'A' will be determinately either true or false. Likewise, the connective 'or' is so understood that 'A' or 'B' is true in those circumstances in which 'A' is true, and in those in which 'B' is true, but in no others. Intuitionistic logic, on the other hand, is founded on the idea that the content of a proposition must be determined by our capacity to recognize it as true or as false. To grasp its content, we must know which recognizable circumstances render it demonstrably true: to assert it is to claim that such circumstances obtain. The meaning of 'or' must be explained by a rule determining the content of 'A or B', as given in this way, from the content of 'A' and of 'B', also so given: namely that those recognizable circumstances which render 'A' demonstrably true, and those which render 'B' demonstrably true, also count as rendering 'A or B' demonstrably true, but that no other recognizable circumstances do so. Since there is plainly no general guarantee, for any proposition 'A', that any recognizable circumstances will obtain that either render 'A', or render 'Not A' demonstrably true, the law of excluded middle 'A or not A' is not, in this logic, a valid law. (Some have seen in the requirement that a proof of 'A or B' constitute a proof either of 'A' or of 'B' an analogy with the legal conception of proof.)

How, then, is the operator 'not' to be understood in this logic? The usual explanation is to pick some patently absurd proposition, say 'Black is white', and explain 'Not A' as equivalent to 'If A, then black is white'. This throws us back on the intuitionistic explanation of 'if', which is that 'If A, then B' is demonstrably true in recognizable circumstances which would render 'B' demonstrably true, given that 'A' were demonstrably true. On the assumption that the absurd proposition 'Black is white' can never be demonstrably true, this amounts to counting 'Not A' as demonstrably true in circumstances which can be recognized as excluding the possibility that 'A' can become recognizably true. The assumption is not built into the logic, however. The fundamental laws governing negation are *reductio ad absurdum*, in the restricted form leading to a negative conclusion, and *ex falso quodlibet*. On the explanation of 'not' in terms of 'if', *reductio ad absurdum* is derivable from the laws governing 'if'. If from the hypothesis 'C' we can infer both 'A' and 'If A, then black is white', only one more step is needed to infer 'Black is white' from 'C': hence, dropping the hypothesis 'C', we may assert, 'If C, then black is white', which is 'Not C'. The *ex falso quodlibet*, however, is not derivable without a further assumption. We need to show that, from the premisses 'A' and 'Not A', we can infer any proposition. Since 'Not A' is 'If A, then black is white', the ordinary laws governing 'if' allow us to infer 'Black is white': so we need to assume that, from the absurd proposition 'Black is white', we can infer any proposition. The stronger assumption, that we shall never be able to assert 'Black is white', here plays no role. If we had a very restricted language, for which there was no absurdity in supposing that every proposition expressible in it might prove to be demonstrably true, it would be sufficient to understand what we have been calling the 'absurd' proposition as the conjunction of all other propositions expressible in the language: all the laws of intuitionistic logic would then hold good. In this sense, then, the intuitionistic meaning of negation – unlike those of the other propositional operators such as 'and' and 'or' – is relative to the language: the weaker the expressive power of the language: the weaker the meaning which logic requires us to impose on the word 'not'.

The word 'contradiction' is often used in a looser sense than the strict one so far discussed. When it is said that there is a contradiction in capitalist economies, it is not meant that the very notion of a capitalist economy involves a contradiction in the strict sense, or otherwise no such thing as a capitalist economy could exist. One of several things may be meant instead. (1) There is a formal contradiction, not in a *description* of a capitalist economy, but in the *justification* normally offered for it, or, perhaps, in any that could be offered. (2) A capitalist economy has a necessary tendency to evolve in each of two incompatible ways, and the resulting social and economic tensions will inevitably destroy it. (3) The term 'capitalist' applies to an economic system as the term 'smooth' applies to a surface, in virtue of its approximation to an unattainable limit: there really is a formal contradiction in the notion of a *pure* capitalist system – one that occupies the limiting position – but not in that of one that approximates to that limit and is hence ordinarily called 'capitalist'. It is in one of these senses, or a similar one, that the word 'contradiction', as customarily employed by Marxists and Hegelians, must be understood; but such a stretched use of it is best avoided. An explicit contradiction, in argument or testimony, is comparatively rare, although of course it does occur: but nothing deserves to be called a contradiction unless it genuinely implies an explicit one.

See Also

- ▶ Axiomatic Theories
- ▶ Existence of General Equilibrium

Contradictions of Capitalism

Andrew Glyn

Writers in the Marxist tradition frequently make use of the term 'contradiction of capitalism'. It is sometimes used, in a very loose sense, to describe virtually any malfunction or indeed objectionable feature of the capitalist system. But in Marx's theory of historical materialism the notion of contradiction played a more fundamental role. One of the central tenets of the theory is that there can be a contradiction between a society's system of economic organization and its capacity to develop its productive potential. Indeed it is precisely such a contradiction between the relations of production (relations of ownership, control etc.) and the forces of production (productive potential), which necessitates through some mechanism or other, a transformation of the economic system. Thus, argued Marx, at a certain stage the rigidities of the feudal system hampered economic growth, which required for its promotion the full and unfettered development of production for the market. The development of productive potential under capitalism formed the basis on which socialism *could* be constructed. The contradictions of capitalism, its inability in turn to take society forward beyond a certain stage, ensured that it *would* be superseded by socialism (see Elster 1985, especially chapter 5).

Labour Power and the Labour Process

For Marx the defining feature of capitalism is that *labour power*, workers' capacity to work, becomes a commodity, which has to be sold by workers who do not have the means of production necessary to work on their own account. The capitalist class pays for this labour power at its value, that is, at a wage determined by social and historical circumstances. But labour power has the capacity to create more value than is contained in it – more precisely, the working class is forced to work longer than is required to produce the goods required to sustain it, leaving a surplus value to be appropriated by the capitalist.

This analysis of the source and nature of profit focuses attention on the factory floor as the locus of the exploitative relation between capital and labour. Labour power is a special commodity in that it cannot be detached from the worker. They do not literally leave their labour power at the factory gate each morning and pick it up in the evening in order to reconstitute it with food and sleep. While this is obvious, it has to be emphasized, since the conventional treatment of production as a matter of technically combining 'labour services' and 'capital services' pays no attention to the active participation of workers in the process of production (see Rowthorn 1980). In fact discipline, supervision, *control* over work are integral to the capitalist system. In turn this means that conflict between workers and employers over all aspects of the labour process is endemic.

Control over labour, and the conflict involved, is clearly a problem for the functioning of capitalism ignored by theories which describe it in terms of the harmonious cooperation between the classes (or owners of factors of production). But does it constitute a contradiction in the sense that it is unresolvable on the basis of private ownership of the means of production, and will lead to increasing malfunctioning of the system as a whole?

It is quite possible to conceive of situations in which inability to control labour in the labour process would become chronic. If it were the case that the development of capitalist production necessarily crowded workers into larger and larger factories, with deteriorating working conditions, but increasing opportunities for organization and resistance, then the question of control over labour could become critical. In fact trends have been more complex. In the advanced capitalist countries, firms have grown enormously in terms of numbers employed, but average plant size has grown much less. Whilst Ford-type production lines may have represented the ultimate in the imposition of capitalist control over the labour process by mechanical means, the continued

requirement for skilled work, demanding judgement, has prevented such systems of work organization being instituted in all industries. Indeed in some industries, worker opposition, or a trend towards more sophisticated products, has led to a reversion to smaller-scale, more integrated methods of production where work is more varied, skilled and responsible.

What is striking, however, is that such trends have in part derived precisely from the resistance engendered by large-scale production. To take the case only of the motor industry, the development of worker resistance in US car plants in the 1960s led to widespread attempts to 'humanize' work by introducing team methods of production and payment. In Italy, conflict in Fiat car factories led to a deliberate policy of decentralizing the less skilled processes of production in order to overcome the problem of controlling 'mass work' in the factories. The production system of Japanese car companies is widely admired, whereby the most important and technically sophisticated stages of production are carried out in large factories, by trained workers, with high wages, paternalistic welfare provisions, tight labour discipline and a modicum of consultation, leaving many components to be produced in much smaller plants by subcontractors, paying lower wages and with less security of employment.

The most important point is a more general one. The shape of development of the capitalist system is determined by the problems and difficulties it encounters. It does not evolve out of some inexorable pattern of technical development; indeed, technology is consciously shaped to overcome social problems (like control over workers) as well as technical ones. A contradiction does not have to spell increasing malfunctioning, let alone capitalism's destruction, to heavily influence the way the system develops.

Labour Shortage

If the first special characteristic of the commodity labour power is that its 'consumption' in the labour process involves the seller (the worker), the second is that its 'production' does not involve the capitalists. For workers are of course 'reproduced' in the home, not produced in factories. The supply of labour power, therefore, cannot like other commodities be increased by a simple redistribution of resources to the sector producing it. The supply of labour, while by no means independent of economic conditions, is not regulated by them as simply as other commodities. Availability of consumer goods does not spell availability of workers. This feature of labour power, together with the issue of control of work already discussed, explains why in analysing production workers cannot be represented by the consumer goods they live on.

The supply of labour is not entirely fixed, of course. Higher wages may increase population growth (as child mortality declines for example), but the social development which accompanies increased living standards may lead to smaller families. This in turn may permit greater participation by women in the labour force. But increased educational standards may delay entry into the labour force, welfare provisions may enable earlier retirement, and part of increased living standards may be taken in reduced hours of work. As pre-capitalist forms of production decline, the possibility for recruiting wage labour from their ranks is diminished; immigration from countries with a labour surplus may meet social and political barriers.

While the supply of labour depends on a host of these factors, not very amenable to short-term manipulation, the demand for labour depends on the rate at which capital is accumulated and its form. Rapid capital accumulation leads to increased demand for labour as workers are required for the new factories. But the new investment may be of a labour-saving variety, requiring fewer workers per machine as compared to earlier vintages. The rise in labour demand depends on the balance between these two forces. If accumulation is sufficiently rapid (as in the advanced capitalist countries in the 1950s and 1960s for example), so that demand for labour rises faster than the supply, then the reserve army of labour (the unemployed and underemployed) shrinks. This improves workers' bargaining position, with consequent difficulties for the employers in

controlling work and wages. A crisis of 'over-accumulation' results.

Increased wages and difficulties in keeping up productivity levels both tend to reduce profits. This leads to reduced investment, insufficient demand for commodities and labour, and stagnation. The 'law of value' does not apply to labour power, so that shortage of supply does not lead to increased profitability in its production and thus increased supply. This can be seen as a fundamental 'contradiction' of capitalism, in the sense that the functioning of capitalism requires labour power to be fully a commodity, and yet this is impossible (see Itoh 1980). Of course this does not establish that the contradiction is irresolvable. If the unemployment which results has the expected effect of reducing workers' bargaining power, then wages can be forced down and productivity up, profits and investment recover and a cyclical upturn results.

Individual and Class Interests

The development of such a crisis of 'over-accumulation' is an example of a more general category of problems. Each individual capitalist is attempting to maximize his profits through securing more labour; yet this leads to lower profits for the capitalist class as a whole as they bid up wages and find increasing problems in work organization. So the rationality of the individual economic agents conflicts with what is rational for the system as a whole. It seems very reasonable to describe this as a 'contradiction' in the functioning of capitalism (Elster 1985). It would require a degree of coordination, which is actually impossible under normal circumstances in a competitive decentralized economy, for the individual employers to hold back from accumulating at a rate which in aggregate is unsustainable. There is no mechanism to tailor the rate of accumulation to what, given the pattern of technical progress, is compatible with the growth of the labour supply, or adjust the pattern of technical progress to what is compatible with the other two variables. What has to 'give' is the rate of profit, and there is no guarantee that the response to a profit squeeze will be a smooth reduction in accumulation to the appropriate level.

There are other examples of 'contradictions' between the interests of individual capitalists and their class interest. Suppose an economic crisis has developed with unused capacity and unemployed labour. Each capitalist may try to improve his competitive position by cutting his employees' wages. But in aggregate the effect of such a strategy would be to reduce consumer demand, which could make the crisis worse. Exactly the same argument applies to the policies of individual capitalist countries trying to solve their problems by increasing their competitiveness. For the context may be a 'negative sum game', whereby cutting wages actually worsens the overall situation. Attempting to cut workers wages, whilst exhorting other capitalists' workers through advertisements to consume more, is a profoundly contradictory situation.

The famous example of this type of contradiction described by Marx was his Law of the Tendency of the Rate of Profit to Fall (LTRPF). He argued that the individual attempts of capitalists to maximize their profits led them to introduce techniques of production which reduced the profit rate for the class as a whole. As described elsewhere (*see* ▶ "Marxist Economics"), Marx's argument is not satisfactory. But this weakness may not seem of great importance, since we have seen in the discussion of overaccumulation that it is perfectly possible to describe a situation where capitalists do act in such a way as to lead to lower profits for them all. The LTRPF leads to a prediction of a continuous decline in the profit rate, and a declining rate of accumulation, leading, if the process developed that far, to absolute stagnation. The actions of capitalists would, in the long run, destroy the very motor of the system, capital accumulation. Crises of overaccumulation, however, are less fundamental in the sense that they are contingent on a particular pattern of accumulation, technical progress and labour supply. Moreover, while they might be repeated there is no basis for asserting an inevitable tendency that they should become deeper and deeper. They can hardly be said therefore to amount to an absolute contradiction in the capitalist process of accumulation, which is the way Marx himself interpreted the LTRPF.

Competition and Concentration

The driving force of capitalism, according to Marx, is competition. This forces the individual capitalist to accumulate capital in the form of new factories, embodying the latest technology. If he fails to do this he will be defeated by his rivals in the battle for markets since his costs will be greater. In modern conditions, where investment is so necessary to generate new products, and where economies of scale in marketing are important alongside those in production, this pressure is stronger than ever. According to Marx the advantages of large-scale production lead to its concentration (he uses the term centralization) in the hands of fewer and fewer firms. As the most dynamic firms knock out, or take over, those that invest less effectively the degree of competition is reduced. At a certain stage this could weaken the pressure to accumulate and generate stagnation in the economy.

Such a contradiction was particularly emphasized by writers basing their ideas on the postwar dominance of giant US firms (see Baran and Sweezy 1966). The development of Japanese and European industry, however, challenged this dominance and, during the 1960s, ushered in a great increase in competition on world markets. While monopolization has increased within each country, there has been a tremendous rise in competition through trade and foreign investment. Some of the Newly Industrialized Countries of South East Asia have begun to break into world markets as well.

The process of competition is, therefore, a complex one. The notion that increased concentration would both reduce the pressure to invest and increase the resources for investment (through higher prices and profits) does not stand as a convincing general trend. That is not to say however that, should a new era of protectionism develop, the high degree of industrial concentration within countries would not exacerbate a tendency to stagnation.

Wasted Resources and Unused Potential

Capitalist production is guided by profit, not social need, or to put it more abstractly, by exchange value rather than use value. The existence of unemployment is the most obvious example of such a contradiction. Unemployed workers could produce the very commodities which they, and the rest of society, need. But since production is for profit, they will only be taken on if the employers foresee a profit. In a situation of unemployment and unused capacity, capital accumulation and thus the introduction of new technology will be held back. The development of technology itself will be reduced if lower profits lead to cuts in research and development spending. For these reasons, society's capacity to produce will be reduced below what is feasible, as well as actual production being reduced below capacity.

These then are some of the senses in which capitalism has been deemed by Marxists to be a 'contradictory' system. The idea, prevalent in the 1950s and 1960s, that these contradictions had been overcome by the expansion of state activities or the advent of the managerial corporation, has disappeared with the collapse of the great postwar boom. Whether capitalism will find a way out of its problems, and lay the basis for rapid growth and full employment, depends of course on how fundamental these contradictions actually are. Even if less binding than some in the Marxist tradition have tended to assert, the idea that such contradictions generate powerful pressures for changes in the economic system remains a powerful and important one.

See Also

- ▶ Economic Interpretation of History
- ▶ Marxist Economics

References

Baran, P., and P. Sweezy. 1966. *Monopoly capital*. New York/London: Monthly Review Press/Penguin.

Elster, J. 1985. *Making sense of Marx*. Cambridge: Cambridge University Press.

Itoh, M. 1980. *Value and crisis*. London: Pluto.

Rowthorn, R.E. 1980. *Capitalism, conflict and inflation*. London: Lawrence & Wishart.

Control and Coordination of Economic Activity

Béla Martos

The particular point of view of the present paper is that it looks upon the economy as a *control system*. This approach was pioneered in the 1950s by Simon (1952), Tustin (1953), Phillips (1954) and Geyer and Oppelt (1957). Lange (1965) attempted an early synthesis. In the 1970s the idea became widespread and developed in two directions. The first and more popular one applied control theoretical models to *economic policy-making*. In this case the structure of the controller is considered to be given and the problem is to find values (time-paths) of the control variables such that the functioning of the economic system be acceptable (most often, stable and/or optimal) according to certain criteria. The second direction is related to the theory of economic systems, and this is where the present paper also belongs. A descriptive and explanatory *theory of economic mechanisms* is aimed at, which might be useful in the choice, change or construction of controllers. Although this research was certainly motivated by, and the findings often applied to, problems emerging in centrally planned economies, with particular reference to mechanism reform in East European countries, the theoretical framework is conceived in a more general setting. This research was initiated by Kornai (1971) and pursued further in Kornai (1980), and Kornai and Martos (1981).

In the first section I present the basic concepts and classifications, followed in the second section by the characterization of the elementary control processes, with the generation and transmission of information and decisions. The final section illustrates the usefulness of this framework by a microeconomic analysis of a non-Walrasian control model.

The Economic Control System

At any point of time (t) an abstract economic system consists of the following ingredients:

A set $\mathcal{A}$ of *agents* (e.g. households, productive firms, banks, government agencies); they are the subjects of the economic activities.

A set $\mathcal{O}$ of *objects* upon which the economic agents act.

The natural, historical, social and economic environment $\mathcal{E}$, which is not a part of, but interacts with, the system.

A set $\ddot{\mathcal{Y}}$ of processes which connect elements of sets $\mathcal{A}$, $\mathcal{O}$ and $\mathcal{E}$ and changes their state.

When speaking about an economic system the first thing we have in mind is a national economy. However, most of the qualifications and methods we use can be applied to systems which are smaller or larger than that (e.g. an industry, a corporation, a region.)

For a consistent control-theory approach two kinds of economic processes (elements of $\ddot{\mathcal{Y}}$) must be distinguished:

Real processes ($\ddot{\mathcal{Y}}_{\mathbf{r}} \subset \ddot{\mathcal{Y}}$), which change the state of physical objects. The most important real processes are production, storage, transfer of physical objects among agents, consumption (whether for productive or for final use). The objects of real processes form the set of *commodities*. ($\mathcal{O}_{\mathbf{r}} \subset \mathcal{O}$), The set of real processes consists of the real activities of the agents and the external effects of the environment. The former ones depend also on the control processes; the external effects cannot be controlled. The rules which connect the real processes are mostly the laws of nature (or more to the point, technology).

Control processes ($\ddot{\mathcal{Y}}_{\mathbf{c}} \subset \ddot{\mathcal{Y}}$), which change the state of knowledge of the agents and regulate their behaviour. The objects of these processes ($\mathcal{O}_{\mathbf{c}} \subset \mathcal{O}$) are called *signals*. The most important control processes are observation of real processes, signal generation and transmission among agents, and decision-making (the final signal generation) on real activities. A part of the signals may come directly from the environment as far as it is observable.

Since each agent $\boldsymbol{a} \in \mathcal{A}$ performs both real and control activities, it is convenient not only to split the set of activities and objects into two (real vs

control) subsets but also to consider each agent as consisting of two units: the *real unit* and the *control unit*, which perform real activities and control, respectively. Needless to say, this splitting of an agent into two units is only a conceptual separation of the functions may correspond with some kind of agents (e.g. large firms), but need not in any organized form exist with other kinds (e.g. households).

Finally, to make the dichotomy of the economic system complete, we can divide even the set $\mathscr{A}$ of agents into two subsets: that of *real agents* (or real organizations), $\mathscr{A}_{\mathbf{r}} \in \mathscr{A}$, whose *main* activities belong to the real processes (like households or productive firms) and that of *control agents* (or control organizations), $\mathscr{A}_{\mathbf{c}} \in \mathscr{A}$, whose *main* activity lies in information-processing and decision-making (like legislative bodies, local authorities, government agencies).

This classification of the agents requires some further comments. Firstly, there might be borderline agents (e.g. schools) whose classification is ambiguous and will be dependent on the role which they play in a given context. Secondly, the real units (real activities) of the control organizations are often negligible in theoretical considerations (just as the energy input of an electric control device might be negligible compared to the energy input of the physical process it controls). We also will make use of this simplification in the sequel and disregard the real activities of the control agents.

Finally, a few words are in order about the place of *fiduciary goods* (banknotes, accounting money, stocks and bonds), *monetary processes* (emission, exchange, income generation, credit) and *financial organizations* (banks, stockbrokers, tax offices) in the above dichotomy. Since it is not the physical transformation of fiduciary goods which is of economic interest (and hence they cannot belong to the real commodities), they belong to the control sphere by exclusion (in contrast with many other theoretical approaches where money is simply taken as one of the commodities). However, it must be kept in mind that the monetary sphere plays not only a particularly important part in the control of economic activities, but is in many aspects different from the rest of the control processes and obeys laws which are partly similar to the ones valid in the real sphere. A thorough discussion of the consequences of this reasoning would require a separate entry.

The economic control system can also be interpreted in the language of mathematical control theory. In a standard state-space representation of a continuously operating, multivariate, deterministic, externally commanded system, it consists of three equations:

Controlled subsystem:

$$\dot{\boldsymbol{x}} = \Phi(\boldsymbol{t}, \boldsymbol{x}, \boldsymbol{u}, \boldsymbol{z}) \tag{1}$$

Measurement:

$$\boldsymbol{y} = \psi(\boldsymbol{x}) \tag{2}$$

Controller:

$$\boldsymbol{u} = \Theta(\boldsymbol{t}, \boldsymbol{y} - \boldsymbol{y}^*), \tag{3}$$

where $t \geq 0$ denotes time and the dot above a variable differentiation with respect to time, $x(t)$ is the state vector, $u(t)$ is the control vector, $y(t)$ is the output vector, $y^*(t)$ is the command vector (the normal value of y), $z(t)$ is the vector of external effects on the state and Φ, Ψ and Θ are functions of their arguments as indicated.

The above system is said to be (globally) *viable* with respect to a closed convex subset $\mathscr{K}$ (the viability set) of the state space (the space of xs) if $\boldsymbol{x} \in \mathscr{K}$ for all $t \geq 0$ and any given initial state $\boldsymbol{x}(\mathbf{0}) = \boldsymbol{x}_{\mathbf{0}} \in \mathscr{K}$. If there is a state $\bar{x} \in \text{Int}\,\mathscr{K}$ and a number $\delta > 0$ such that $\boldsymbol{x} \in \mathscr{K}$ for all $t \geq 0$ and any given initial state $x_o \in \mathscr{K} \cap \{x \mid \|x - \bar{x}\| < \delta\}$, that is in the neighbourhood of $\bar{x}$, then the system is said to be *locally viable* at $\bar{x}$ with respect to $\mathscr{K}$.

It was proved by Aubin and Cellina (1983, theorem 5.4.1) that under some continuity, convexity and compactness assumptions there is a feedback rule Θ such that the system (1) to (3) is globally viable. It is to be noted, however, that this is an existence theorem from which no conclusion

can be drawn, in this generality, as to how the appropriate feedback rule Θ can be constructed.

The form (1) to (3) is, of course, not the only mathematical form in which a control system can be represented, but it is general enough to cover many important cases and special forms, which are too numerous to list here even partially. I would rather mention systems which are not explicitly represented by the above formulation.

(a) *Intermittently operating systems*. It is frequently the case that, especially in economic applications, the measurement of the state is not done continuously but only at discrete points of time. In this case the value of the control variable remains constant in between. If the observation times are equidistant, the above formulation can easily be transformed to cover this case simply by replacing the differential operator of the left-hand side by a time shift operator $Ex(t) = x(t+1)$

(b) *Stochastic systems* arise if x and/or y and/or u represent stochastic processes, and consequently some of the operators, Φ, Ψ Θ have stochastic values. In the case of a stochastic Φ, the controlled system works erratically; a stochastic ψ indicates measurement errors; and a stochastic Θ indicates uncertain control behaviour. These are frequent cases in economic systems. (It is to be noted that any random disturbance on z and y^*, i.e. on variables representing the environment, does not make the system stochastic, they are the realizations which enter the functions.)

(c) *Optimum control*, in which case the control rule is not given in the form (3) but is rather a solution to the problem of maximizing a given functional

$$\mathrm{I} = \int_0^T \mu(t, x, u)\mathrm{d}t$$

subject to (1) and some other constraints which require the control variable u to belong to a given set $\mathscr{U}$, and where μ is a scalar function of the arguments.

(d) *Higher-order systems* (as contrasted to externally commanded systems) take different forms:

Self-command (or target modifying) systems produce the command signals y^* themselves.

Learning systems modify the form or parameter values characterizing the operator Θ ; a learning mechanism improves the controller.

Self-organizing systems are capable of changing the control structures, the organizations and the interrelations among them both in the controlled subsystem and the controller.

Although it is clear that most economic systems perform such higher-order functioning, their mathematical analysis is difficult and mostly reduced to narrowly specified cases.

The Structure of the Controller

The controller was typified in equation (3) in a very rough-and-ready way. In actual economic systems the controller has a rather complicated structure, consisting of many different elements which interact in various ways. Some of the elements make simple observations, routine calculations, bookkeeping, and so on; others collect, generate and disseminate important information or make crucial decisions and plans relying on a vast amount of preprocessed information. Some of them work in parallel on different sets of data, and some form interactive or hierarchically ordered groups.

The study of such a structure must begin with the functioning of its constituent elements which are called *transfer elements*. A transfer element is an elementary part of a complex controller which cannot be divided further or has not been in a particular analysis.

There are three subsequent actions in the functioning of a transfer element:

Signal reception. The transfer element receives signals (information) from the observation of real processes, from the environment or from another transfer element. These are the *input signals* of the element.

Signal transformation or signal generation. The transfer element transforms, stores and combines the received signals and hereby generates new ones. The rules by which signals are generated form the *transfer function* of the element.

Signal emission. The transfer element transmits the generated signal (*output signal*) to one or more other transfer elements or to an agent which acts directly on real processes.

In the classification of the elementary control process we apply two criteria both with respect to the kinds of agents who participate in the process: – What kind of agent generates the signal? – Among what kind of agents is the signal transmitted?

With respect to *signal generation* we distinguish three kinds of processes:

Uncoordinated. The signal is generated by the control unit of a single real organization.

Interactive. The signal is generated jointly by the control units of several real organizations.

Centralized. The signal is generated by a control organization or jointly by several control and perhaps real organizations.

With respect to *signal transmission* we also distinguish three kinds of process:

Non-communicative. The signal does not leave the organization where it was generated.

Transactional. The sender and the addressee are two different real organizations, and the signal refers to an (actual or potential) real transaction (usually transfer of a commodity) between the two real organizations (e.g. dispatch of an order, a price quotation, a bill).

Communicative (non-transactional). Any other signal transmission; for example, among more than two real organizations, or whenever a control organization is the sender or the addressee or both.

This dual classification of the transfer elements can be summarized in below table. The two empty boxes represent signal generation–transmission combinations which cannot occur. (An interactive signal generation implies some kind of communication, since to generate signals jointly by several real organizations, they must communicate somehow. In the centralized signal generation a control organization takes part, hence it cannot be transactional.)

This simple classification scheme can be applied to elementary transfer units of the controller only. In a complex control process several transfer units are combined which differ with respect to their signal generation and transmission patterns.

	Signal generation		
	Uncoordinated	Coordinated	
Signal transmission		Interactive	Centralized
Non-communicative	+	Ø	+
Transactional	+	+	Ø
Communicative (non-transactional)	+	+	+

Most of the actually existing economic control systems may be called *partially coordinated systems*, in which a considerable part of the decisions are taken by the real organizations in isolation, another part by their interaction (e.g. on the market) and yet another part by different control agents (e.g. legislative bodies, government agencies, banks, trade unions etc.). The problem of analysing (synthesizing) an economic control system consists of the decision about whether one or the other function of the system is (should be) served by this or that kind of transfer unit and how these units are (can be) integrated into a viable or even efficient entity.

An essential feature of the above conceptualization of the structure of the economic control system is that it does not restrict the issue to 'control and coordination of economic activity' from the outside (done exclusively by specialized control organizations) but includes the control functions which work within the real organizations and interact among them. It is also to be noted that the classical distinction between centralized and decentralized control turned out to be insufficient and has been replaced by a more elaborate classification pattern.

A Non-Walrasian Control Structure

The first economic theory which offered a mathematically rigorous representation of the control mechanism of a national economy is known under the term *General Equilibrium Theory.* Since neither Keynesian macroeconomics in capitalistic systems nor shortage phenomena in socialist economies could have been appropriately studied within the framework of this theory, a new approach emerged under various names: *disequilibrium theory, temporary equilibrium theory, theory of equilibria with rationing, non-Walrasian equilibrium theory.* Without discussing here merits and demerits of these approaches, it is to be noted that – as a rule–they were not based on mathematical control theory.

In what follows I present a non-Walrasian control model differing from the aforementioned approaches in many aspects:

(a) It is not only the (essentially static) equilibrium, its existence and efficiency which is studied, but rather the dynamics of the trajectories leading to an equilibrium state. Real and control processes run in parallel (out of equilibrium); there is no timeless *tâtonnement* process.
(b) No optimizing behaviour of the agents is assumed; adjustment to exogenous normal values of some output variables is the behavioural rule. When applying this 'control by norm' principle I assume that norms are formed by individual experience or social consent in a long-run process (which is not modelled here), and the norms remain constant along the short-and medium-run adjustment process.
(c) Information and decisions are not centralized as in the hands of an auctioning or rationing agent, but the whole control process is carried out by the control units of real organizations among themselves in an uncoordinated but transactionally communicative way. (This refers only to the particular model variant which follows. In other variants control organizations and coordination also appear.)
(d) Only observable variables are used (no fictitious 'effective demand') and hence the underlying assumptions can be, but generally have not been, empirically tested. (For an exception, see Kawasaki et al. 1982.)

Still it is to be admitted that this approach has not yet reached the generality and mathematical refinement of general equilibrium and disequilibrium theory.

The model. The economy consists of n producers (real organizations), each producing a single commodity. The technology is of the Leontief-type, with constant input coefficients. The environment acts upon the real processes by the final use (private and public consumption, investment) and on the control processes by past experiences, which determine the normal level of inventories (output stocks, input stocks) and backlog orders.

Notation: lower case–n-vector; upper case–$n \times n$ matrix; Greek lower case–scalar.

State variables:

q – vector of output stocks
V – matrix of input stocks
K – matrix of backlog orders

An asterisk * as a superscript refers to the exogenous normal values of the state variables.

Control variables:

r – vector of production ($\langle r \rangle$: the diagonal matrix formed from r)
Y – matrix of commodity transfers among producers
W – matrix of the transmission of new orders

Other notations:

$e = [1,1,\ldots 1]'$ – the summation vector
A – the input coefficient matrix
c – the vector of final uses
β, γ – control parameters
$\Gamma(\cdot) = -2\beta\gamma\,[d(\cdot)/\mathrm{dt}] - \gamma^2 \cdot (\cdot)$-differential operator.

Assumptions:

1. The final use is constant and semipositive, $c \geq 0$.

2. The input coefficient matrix A is constant and
   (a) non-negative
   (b) irreducible
   (c) productive, i.e. its spectral radius $\rho(A) < 1$.
3. $\gamma > 0$ (without loss of generality).

The real processes:

$$\dot{q} = r = Ye - c \tag{4}$$

$$\dot{V} = Y - A\langle r \rangle. \tag{5}$$

Equation (4) expresses the change of output stocks as the difference between the amounts produced and that transferred for productive and final use. Equation (5) tells that the change of input stocks equals the material purchases minus the materials used up in production.
The control processes:

$$\dot{K} = W - Y \tag{6}$$

$$\dot{\boldsymbol{r}} = \Gamma(\boldsymbol{q} - \boldsymbol{q}^*) \tag{7}$$

$$\dot{\boldsymbol{W}} = \boldsymbol{\Gamma}(\boldsymbol{V} - \boldsymbol{V}^*) \tag{8}$$

$$\dot{\boldsymbol{Y}} = -\boldsymbol{\Gamma}(\boldsymbol{K} - \boldsymbol{K}^*). \tag{9}$$

Equation (6) describes the bookkeeping (at the supplier) of the backlog of orders; its change equals the difference between the incoming new orders and the deliveries. Equations (7) to (9) are the control equations proper, all of the same (linear) form, describing the assumed behaviour of the agents. The decisions on production level is dependent on the output stocks, the dispatch of orders (by the buyer) on the input stocks, and the deliveries (decided by the supplier) on the backlog of orders, in each case taking the deviation of the actual value from the normal value into account. None of these behavioural rules is at variance with common sense.

It is to be observed, that the transfer elements corresponding to equations (6) to (9) generate all the signals without any coordination; equations (6), (7) and (9) represent non-communicative elements, while there is transactional communication according to equation (8); namely, the orders are transmitted from the buyers to the suppliers.

The viability domain $\mathscr{K}$ for system (4) to (9) may be defined in the following way:

(a) All the variables are uniformly bounded, but the bounds are unspecified.
(b) The variables in q, V, K, r and Y are non-negative, but negative elements of W (withdrawal of orders) are permitted.

Although the theorem of Aubin and Cellina referred to above does not apply here, where was specified the form of the control equations (6) to (9), we can still guarantee local viability in the neighbourhood of the equilibrium state by an appropriate choice of the parameter β.

Theorem. Suppose that the following conditions are met:

(a) Assumptions (4) to (6) hold.
(b) The norms are positive: $q^*>0$, $V^*>0$, $K^*>0$.
(c) $\beta > \max\{|\operatorname{Im}\sigma| / (2|\sigma|)\sqrt{\operatorname{Re}\sigma|} - \sigma^3 + 2\sigma^2 - 2\sigma + 1 \in \text{spectrum of } A\}$ and $\beta > \sqrt{6/4}$.
(d) The initial values at $t = 0$: $(q^0, V^0, K^0, r^0, Y^0, W^0)$ are close enough to the equilibrium state:

$$\overline{q} = q^*, \quad \overline{V} = V^*, \overline{K} = K^*,$$
$$\overline{r} = (E - A)^{-1}c, \quad \overline{Y} = \overline{W} = A\left\langle (E - A)^{-1}c \right\rangle.$$

Then the system (4) to (9) is viable for $t \geq 0$ (local viability).

Remark: under (a) the relation (c) is both a necessary and sufficient condition of asymptotic stability.

A detailed analysis of the model and proof of the theorem (extended to varying c) is to be found in a forthcoming book by Martos. Models in a similar vein are analysed in Kornai and Martos (1981).

See Also

- ▶ Decentralization
- ▶ Planned Economy
- ▶ Planning
- ▶ Pontryagin's Principle of Optimality

References

Aubin, J.P., and A. Cellina. 1983. *Differential inclusions*. Berlin: Springer.
Geyer, W., and W. Oppelt, ed. 1957. *Volkswirtschaftliche Regelungsvorgänge im Vergleich zu Regelungsvorgängen in der Technik*. Munich.
Kawasaki, S., J. McMillan, and K.F. Zimmermann. 1982. Disequilibrium dynamics: An empirical study. *American Economic Review* 72 : 992–1003.
Kornai, J. 1971. *Anti-equilibrium*. Amsterdam: North-Holland.
Kornai, J. 1980. *Economics of shortage*. Amsterdam: North-Holland.
Kornai, J., and B. Martos, ed. 1981. *Non-price control*. Amsterdam: North-Holland.
Lange, O. 1965. *Wstep do cybernetyki ekonomicznej* (Introduction to economic cybernetics). Warsaw: Państwowe Wydawnictwo Naukowe.
Phillips, A.W. 1954. Stabilization policy in a closed economy. *Economic Journal* 64 : 290–323.
Simon, H.A. 1952. On the application of servomechanism theory in the study of production control. *Econometrica* 20 : 247–268.
Tustin, A. 1953. *The mechanism of economic systems*. London: Heinemann.

Control Functions

Salvador Navarro

Abstract

The control function approach is an econometric method used to correct for biases that arise as a consequence of selection and/or endogeneity. It is the leading approach for dealing with selection bias in the correlated random coefficients model. The basic idea of the method is to model the dependence between the variables not observed by the analyst on the observables in a way that allows us to construct a function K such that, conditional on the function, the endogeneity problem (relative to the object of interest) disappears.

Keywords

Average treatment effect; Control functions; Endogeneity; Identification; Instrumental variables; Roy model; Selection bias

JEL Classifications

C1

The control function approach is an econometric method used to correct for biases that arise as a consequence of selection and/or endogeneity. It is the leading approach for dealing with selection bias in the correlated random coefficients model (see Heckman and Robb 1985, 1986; Heckman and Vytlacil 1998; Wooldridge 1997, 2003; Heckman and Navarro 2004), but it can be applied in more general semiparametric settings (see Newey et al. 1999; Altonji and Matzkin 2005; Chesher 2003; Imbens and Newey 2006; Florens et al. 2007).

The basic idea behind the control function methodology is to model the dependence between the variables not observed by the analyst on the observables in a way that allows us to construct a function K such that, conditional on the function, the endogeneity problem (relative to the object of interest) disappears.

In this article I deal exclusively with the problem of identification. That is, I assume access to data on an arbitrarily large population. As a consequence, I do not discuss estimation, standard errors or inference. In the examples, I analyse how to recover parameters in a way that, I hope, shows directly how to perform estimation via sample analogues.

The Set-Up

The general set-up I consider is the following two-equation structural model; an outcome equation:

$$Y = g(x, D, \varepsilon), \tag{1}$$

and an equation describing the mechanism assigning values of D to individuals:

$$D = h(X, Z, v), \tag{2}$$

where X and Z are vectors of observed random variables, D is a (possibly vector valued) observed random variable, and ε and v are general disturbance vectors not independent of each other but satisfying some form of independence of X and Z.

The problem of endogeneity arises because D is correlated with ε via the dependence between ε and v. Because Eq. (2) represents an assignment mechanism in many economic models, it is generically called the 'selection' or 'choice' equation. This set-up has been applied to problems like earnings and schooling (Willis and Rosen 1979; Cunha et al. 2005), wages and sectoral choice (Heckman and Sedlacek 1985) and production functions and productivity (Olley and Pakes 1996), among others.

The goal of the analysis is to recover some functional of $g(X, D, \varepsilon)$ of interest

$$a(X, D) \tag{3}$$

that cannot be recovered in a straightforward way because of the endogeneity/ selection problem. As an example, when D is binary interest sometimes centres on the effect of going from $D = 0$ to $D = 1$ for an individual chosen at random from the population, the so-called average treatment effect:

$$a(X, D) = E(g(X, 1, \varepsilon) - g(X, 0, \varepsilon)).$$

The key behind the control function approach is to notice that (conditional on X, Z) the only source of dependence is given by the relation between ε and v. If v was known, we could condition on it and analyse Eq. (1) without having to worry about endogeneity. The main idea behind the control function approach is to recover some function of v via its relationship with the model observables so that we can now condition on it and solve the endogeneity problem.

Definition The control function approach proposes a function K (the control function) that allows us to recover a (X, D) such that K satisfies

A-1. K is a function of X, Z, D.

A-2. ε satisfies some form of independence of D conditional on ρ (X, K), with ρ a knowable function.

A-3. K is identified.

Assumption **A-2** is the key assumption of the approach. It states that, once we condition on K, the dependence between ε and D (that is, the endogeneity) is no longer a problem. To help fix ideas, consider the following example of a simple linear in parameters additively separable version of the model of Eqs. (1 and 2).

Example 1 *Linear regression with constant effects. Write the outcome Eq. (1) as*

$$Y = X\beta + D\alpha + \varepsilon$$

and assume that our object of interest (3) is α. Assume that we can write Eq. (2) as

$$D = X\rho + Z\pi + v \tag{4}$$

with v, $\varepsilon \perp\!\!\!\perp X$, Z where $\perp\!\!\!\perp$ denotes statistical independence. Such a model arises, for example, if Y is logearnings and D is years of schooling as in Heckman et al. (2003). If ability is unobservable since high ability is associated with higher earnings but also with higher schooling, then ε and v would be correlated.

If we let $K = v$ be the residual of the regression in (4), then we can recover a from the following regression

$$Y = X\beta + D\alpha + K\psi + \eta,$$

where it follows that $E(\eta|X, K) = 0$. It is easy to show that in this case the control function estimator and the two-stage least squares estimator are equivalent. (To my knowledge, although in a different context – a SUR model – Telser 1964, was the first to use the residuals from other equations as regressors in the equation of interest.)

The previous case is a simple example of a control function where $K = D - E(D|X, Z)$. In

this case, because of the constant effects assumption (that is, α is not random), standard instrumental variables methods and the control function approach coincide. In general, this is not the case.

In the next section I describe in detail the control function methodology for the binary choice case (Roy 1951). This case is interesting both because it is the workhorse of the policy evaluation literature and because, by virtue of its nonlinearity, it highlights the implications of a nonlinear structure in a relatively simple context. I then briefly describe extensions to more general cases. For simplicity, I focus on the additively separable in unobservables case, but recent research provides generalizations to non-additive functions (see Blundell and Powell 2003; Imbens and Newey 2006, among others).

The Case of a Binary Endogenous Variable

In this section I describe how the control function approach solves the selection/ endogeneity problem when the endogenous variable is binary. This problem has a long tradition in economics going back (at least) to Roy (1951). In Roy's original version of the model (see Roy model) an individual is deciding whether to become a fisherman ($D = 0$) or a hunter ($D = 1$).

Associated with each occupation is a payoff $Y_D = g_D(X) + \varepsilon_D$. Since we can only observe individuals in one sector at a time, the *observed* outcome for an individual is given by Y_1 if he becomes a hunter ($D = 1$) and by Y_0 if he becomes a fisherman ($D = 0$). That is, the observed outcome (Y) can be written as:

$$\begin{aligned} Y &= DY_1 + (1-D)Y_0 \\ &= g_0(X) + D(g_1(X) - g_0(X)) + \varepsilon_0 + D(\varepsilon_1 - \varepsilon_0). \end{aligned} \tag{5}$$

The model is closed by assuming that individuals choose the occupation with the highest payoff. That is,

$$\begin{aligned} D &= 1(Y_1 - Y_0 > 0) \\ &= 1(g_1(X) - g_0(X) + \varepsilon_1 - \varepsilon_0 > 0), \end{aligned} \tag{6}$$

where $\mathbf{1}(a)$ is an indicator function that takes value 1 if a is true and 0 if it is false. Endogeneity arises because the error term in choice Eq. (6) contains the same random variables as the outcome Eq. (5). A generalized version of the model replaces the simple income maximization rule in (6) with a more general decision rule

$$D = 1(h(X,Z) - v > 0). \tag{7}$$

The model described by Eqs. (5 and 7) is general enough to be used in many different cases. Many qsts of interest in economics fit this framework if, instead of thinking of two sectors, fishing and hunting, we think of two generic potential states, the treated state ($D = 1$) and the untreated state ($D = 0$) with their associated potential outcomes. The decision rule in (7) is general enough to capture not only income maximization but also utility maximization and even a deciding actor different from the agent directly affected by the outcomes (parents deciding for their children, for example). The simple income maximization rule in (6) shows why, in *general* if $\varepsilon_1 \neq \varepsilon_0$, then $\varepsilon_1 - \varepsilon_0$ is likely to be correlated with D.

The correlated random coefficients model is a special case of the model described by (5) and (7) when $\varepsilon_1 - \varepsilon_0$ is not independent of D and $g_j(X) = \alpha_j + X\beta$ for $j = 0,1$. (For simplicity I assume $\beta_1 = \beta_0 = \beta$. The case where $\beta_1 \neq \beta_0$ follows directly.) To see why simply rewrite (5) as

$$Y = \alpha_0 + X\beta + D(\alpha_1 - \alpha_0 + \varepsilon_1 - \varepsilon_0) + \varepsilon_0 \tag{8}$$

so that now the coefficient on D is (a) random and (b) correlated with D. In this case we have that the gains from treatment ($\alpha_1 - \alpha_0 + \varepsilon_1 - \varepsilon_0$) are heterogeneous (that is, they are not constant even after controlling for X) and they are correlated with D. I come back to this special linear in parameters case in Example 2.

Though other parameters of interest can be defined, I consider the case in which we are interested in the two particular functionals that receive the most attention in the evaluation literature – the average treatment effect and the average effect of treatment on the treated. I impose that ε_1, ε_0, v are absolutely continuous with finite means, and that

$\varepsilon_1, \varepsilon_0, v \perp\!\!\!\perp X, Z$. (One could weaken the assumption to be $\varepsilon_1, \varepsilon_0 \perp\!\!\!\perp X|Z$ and $v \perp\!\!\!\perp X, Z$.)

Under these assumptions the average treatment effect is given by

$$\begin{aligned} ATE(x) &= E(Y_1 - Y_0 | X = x) = g_1(x) - g_0(x) \\ &= x(\beta_1 - \beta_0) \end{aligned}$$

where the last equality follows if Eq. (8) applies. $ATE(X)$ is of interest to answer qsts like the average effect of a policy that is mandatory, for example. When receipt of treatment is not mandatory or randomly assigned, the average effect of treatment among those individuals who are selected into treatment is commonly the functional of interest (see Heckman 1997; Heckman and Smith 1998). This effect is measured by the average effect of treatment on the treated:

$$\begin{aligned} TT(x) &= E(Y_1 - Y_0 | X = x, D = 1) \\ &= g_1(x) - g_0(x) + E(\varepsilon_1 - \varepsilon_0 | X = x, D = 1) \\ &= \alpha_1 - \alpha_0 + E(\varepsilon_1 - \varepsilon_0 | X = x, D = 1), \end{aligned}$$

where the last equality follows for the linear in parameters case of Eq. (8).

Now, suppose we ignored the endogeneity problem and attempted to recover either of these objects from the data on outcomes at hand. In particular, if we used the (observed) conditional means of the outcome

$$\begin{aligned} E(Y | X = x, D = 1) - E(Y | X = x; D = 0) \\ = g_1(x) - g_0(x) + E(\varepsilon_1 | X = x, D = 1) \\ - E(\varepsilon_0 | X = x, D = 0) \end{aligned}$$

we would not recover either $ATE(X)$ or $TT(x)$. Notice too that, since the endogenous variable D is binary, we cannot directly recover v and use it as a control as we did in the linear case of Example 1 above. Instead, we can recover a function of v that satisfies the definition of a control function.

Let $F_v()$ denote the cumulative distribution function of v. To form the control function in this case, first take Eq. (7) and write the choice probability

$$\begin{aligned} P(x, z) &= \Pr(D = 1 | X = x, Z = z) \\ &= \Pr(v < h(x, z)) = F_v(h(x, z)), \end{aligned}$$

which under our assumptions implies

$$h(x, z) = F_v^{-1}(P(x, z)).$$

Following the analysis in Matzkin (1992), we can recover both $h(x, z)$ and $F_v()$ nonparametrically up to normalization.

Next, take the conditional (on X, Z) expectation of the outcome for the treated group

$$\begin{aligned} E(Y | X = x, Z = z, D = 1) \\ = g_1(x) + E(\varepsilon_1 | X = x, Z = z, D = 1). \end{aligned}$$

We can write the last term as

$$\begin{aligned} E(\varepsilon_1 | X = x, Z = z, D = 1) &= E(\varepsilon_1 | v < h(x, z)) \\ &= E\left(\varepsilon_1 | v < F_v^{-1}(P(x, z))\right). \end{aligned}$$

That is, we can write it as a function of the known h(x, z) or, equivalently, as a function of the probability of selection $P(x, z)$,

$$\begin{aligned} E(Y | X = x, Z = z, D = 1) \\ = g_1(x) + K_1(P(x, z)), \end{aligned}$$

where $K_1(P(X, Z))$ satisfies our definition of a control function. So, provided that we can vary $K_1(P(X, Z))$ independently of $g_1(X)$, we can recover $g_1(X)$ up to a constant. We can identify the constant in a limit set such that P → 1 since $\lim_{P \to 1} K_1(P) = 0$. Provided that we have enough support in the probability of treatment – that is, provided that some people choose treatment with probability arbitrarily close to (1) – we can recover the constant. (See Example 2.) Using the same argument we can form

$$\begin{aligned} E(Y | X = x, Z = z, D = 0) \\ = g_0(x) + K_0(P(x, z)) \end{aligned}$$

and identify $g_0(X)$ (up to a constant) and the control function K_0(P(X, Z)). As before, we can recover the constant in $g_0(X)$ by noting that $\lim_{P \to 0} K_0(P) = 0$.

Intuitively, we need to be able to vary the $K_1(P(X, Z))$ function relative to the $g_1(X)$ function

so that we can identify them from the observed variation in Y_1. One possibility is to impose that g_1 and K_1 are measurably separated functions. (That is, provided that, if $g_1(X) = K_1(P(X, Z))$ almost surely then $g_1(X)$ is a constant almost surely; see Florens et al. 1990.) The simplest way to satisfy this restriction is by exclusion. That is, if $K_1(P(X, Z))$ is a nontrivial function of Z conditional on X and Z shows enough variation, we can vary the K_1 function by varying Z while keeping g_1(X) constant. Another related possibility is to assume that g_1 and K_1 live in different function spaces. For example, g_1 a linear function and K_1 the nonlinear mills ratio term that results from assuming that (ε_0, ε_1, v) are jointly normal as in the original Heckman (1979) selection correction model.

Once we have recovered $g_0(X)$, $g_1(X)$, K_0 ($P(X, Z)$), K_1 ($P(X, Z)$) we can now form our parameters of interest. Given $g_0(X)$ and $g_1(X)$, $ATE(X) = g_1(X) - g_0(X)$ immediately follows. To recover $TT(X)$, first notice that, by the law of iterated expectations

$$\begin{aligned} E(\varepsilon_0|X = X, Z = z) &= E(\varepsilon_0|X = x, Z = z, D = 1)P(x, z) \\ &\quad + E(\varepsilon_0|X = x, Z = z, D = 0)(1 - P(x, z)) \\ &= 0, \end{aligned}$$

where P(X, Z) is known from our analysis above and $E(\varepsilon_0|\ X = x,\ Z = z,\ D = 0) = K_0(P(X, z))$. Rewriting the expression above we get $E(\varepsilon_0|X = x, Z = z, D = 1) = \frac{K_0(P(x,z))(1-P(x,z))}{P(x,z)}$. With this expectation in hand we can recover $TT(X, Z) = g_1(X) - g_0(X) + K_1(P(X, Z)) + \frac{K_0(P(X,Z)(1-P(X,Z))}{P(X,Z)}$. By integrating against the appropriate distribution, we can recover $TT(X) = \int TT(X, z)dF_{Z|X,\ D\ =\ 1}(z)$.

The following example shows how the control function methodology can be applied to recover average effects of treatment in a linear in parameters model with correlated random coefficients. This model arises when there are unobservable gains that vary over individuals and these gains are correlated with the choice of treatment (that is, when there is essential heterogeneity. See Heckman et al. 2006; Basu et al. 2006). The Roy model of Eqs. (5 and 6) in which the unobservable individual gains ($\varepsilon_1 - \varepsilon_0$) are correlated with the choice of sector is an example of this case.

Example 2 *Correlated random coefficients with binary treatment.* Assume we can write the outcome equations in linear in parameters form,

$$Y_j = \alpha_j + X\beta_j + \varepsilon_j j = 0, 1.$$

Let D be an indicator of whether an individual receives treatment ($D = 1$) or not ($D = 0$). We also write a linear in parameters decision rule:

$$D = 1(X\delta + Z\gamma - v > 0).$$

From the analysis in Manski (1988) we can recover δ, γ and F_v (up to scale). With $P(x, z) = \Pr(D = 1|X = x, Z = z)$ in hand, we then form

$$Y_j = \alpha_j + X\beta_j + K_j(P(X, Z)) + \eta_j$$

where $E(\eta_j|\ X = x,\ K_j(P(X, Z)) = k_j) = 0$. To emphasize the problem of identification of the constant α_j we can rewrite the outcome as

$$Y_j = \tau_j + X\beta_j + \tilde{K}_j(P(X, Z)) + \eta_j$$

where $K_j = (P(X, Z)) = \kappa_j + \tilde{K}_j(P(X, Z))$ and $\tau_j = \alpha_j + \kappa_j$.

The elements of the outcome equations can be recovered by various methods. One could, for example, use Robinson (1988) and use residualized nonparametric regressions to recover β_j, τ_j and $K_j(P(X, Z))$. Alternatively, one could approximate $K(P(X,\ Z))$ with a polynomial on $P(X, Z)$. In this case we would have

$$\begin{aligned} Y_j = \tau_j + X\beta_j + \pi_1 P(X, Z) + \pi_2 P(X, Z)^2 + \cdots \\ + \pi_n P(X, Z)^n + \eta_j \end{aligned}$$

where $\tilde{K}_j(P(X, Z)) = \sum_{i=1}^{n} \pi_{j1} P(X,\ Z)^i$. When $j = 0$ then $\lim_{P \to 0} K_0(P) = 0$ and it follows that $\tilde{K}_0(P) = K_0(P)$ and $\tau_0 = \alpha_0$. For the treated case ($j = 1$) we have that $lim_{P \to 1} K_1(P(X, Z)) = 0$. Since $\tilde{K}_1(1) = \sum_{i=1}^{n} \pi_{1i}$ it follows that $\kappa_1 = -\sum_{i=1}^{n} \pi_{1i}$ and $\alpha_1 = \tau_1 - \sum_{i=1}^{n} \pi_{1i}$.

Extensions for a Continuous Endogenous Variable

In this section I briefly review the use of the control function approach for the case in which the endogenous variable D is continuous and we assume that $X, Z \perp\!\!\!\perp \varepsilon, v$. Following Blundell and Powell (2003) I assume that the object of interest is the average structural function

$$a(X,D) = \int g(X,D,\varepsilon) dF_{\varepsilon}(\varepsilon),$$

which, in the additively separable case $g(X, D, \varepsilon) = \mu(X, D) + \varepsilon$ is simply the regression function $\mu(X, D)$.

If we assume that the choice equation

$$D = h(X,Z,v)$$

is strictly monotonic in v (which would follow automatically if it were additively separable in v), we can recover $h()$ and F_v from the analysis of Matzkin (2003) up to normalization. A convenient normalization is to assume that $v \sim$ Uniform (0,1) in which case we can directly recover v from the quantiles of F_v, but other normalizations are possible. From the independence assumption it follows that $E(\varepsilon|X, D, Z) = E(\varepsilon|v)$, so we can write the outcome equation as

$$\begin{aligned} Y &= \mu(X,D) + E(\varepsilon|\, v) \\ &= \mu(X,D) + K(v) \end{aligned}$$

which allows us to recover $\mu(X, D)$ directly (up to normalization). In the additively separable case we analyse, we can relax the full independence assumption and instead assume directly that the weaker mean independence assumption $E(\varepsilon|X, D, Z) = E(\varepsilon|v)$ holds.

See Also

- ▶ Endogeneity and Exogeneity
- ▶ Identification
- ▶ Roy Model
- ▶ Selection Bias and Self-Selection

Bibliography

Altonji, J.G., and R.L. Matzkin. 2005. Cross section and panel data estimators for nonseparable models with endogenous regressors. *Econometrica* 73: 1053–1102.

Basu, A., J.J. Heckman, S. Navarro, and S. Urzua. 2006. Use of instrumental variables in the presence of heterogeneity and self-selection: An application in breast cancer patients. Unpublished manuscript, Department of Medicine, University of Chicago.

Blundell, R., and J. Powell. 2003. Endogeneity in nonparametric and semiparametric regression models. In *Advances in economics and econometrics: Theory and applications, eighth world congress*, ed. L.P. Hansen, M. Dewatripont, and S.J. Turnovsky, Vol. 2. Cambridge: Cambridge University Press.

Chesher, A. 2003. Identification in nonseparable models. *Econometrica* 71: 1405–1441.

Cunha, F., J.J. Heckman, and S. Navarro. 2005. Separating uncertainty from heterogeneity in life cycle earnings. *Oxford Economic Papers* 57: 191–261.

Florens, J.-P., M. Mouchart, and J.M. Rolin. 1990. *Elements of Bayesian statistics*. New York: M. Dekker.

Florens, J.-P., J.J. Heckman, C. Meghir, and E.J. Vytlacil. 2007. Identification of treatment effects using control functions in models with continuous, endogenous treatment and heterogeneous effects. Unpublished manuscript, Columbia University.

Heckman, J.J. 1979. Sample selection bias as a specification error. *Econometrica* 47: 153–162.

Heckman, J.J. 1997. Instrumental variables: A study of implicit behavioral assumptions used in making program evaluations. *Journal of Human Resources* 32: 441–462. Addendum published in 33(1) (1998).

Heckman, J.J., and S. Navarro. 2004. Using matching, instrumental variables, and control functions to estimate economic choice models. *Review of Economics and Statistics* 86: 30–57.

Heckman, J.J., and R. Robb. 1985. Alternative methods for evaluating the impact of interventions: An overview. *Journal of Econometrics* 30: 239–267.

Heckman, J.J., and R. Robb. 1986. Alternative methods for solving the problem of selection bias in evaluating the impact of treatments on outcomes. In *Drawing inferences from self-selected samples,* ed. H. Wainer. New York: Springer. Repr. Mahwah: Lawrence Erlbaum Associates, 2000.

Heckman, J.J., and G.L. Sedlacek. 1985. Heterogeneity, aggregation, and market wage functions: An empirical model of self-selection in the labor market. *Journal of Political Economy* 93: 1077–1125.

Heckman, J.J., and J.A. Smith. 1998. Evaluating the welfare state. In *Econometrics and economic theory in the twentieth century: The ragnar frisch centennial symposium*, ed. S. Strom. New York: Cambridge University Press.

Heckman, J.J., and E.J. Vytlacil. 1998. Instrumental variables methods for the correlated random coefficient model: Estimating the average rate of return to

schooling when the return is correlated with schooling. *Journal of Human Resources* 33: 974–987.
Heckman, J.J., L.J. Lochner, and P.E. Todd. 2003. Fifty years of mincer earnings regressions. Technical Report No. 9732. Cambridge, MA: NBER.
Heckman, J.J., S. Urzua, and E.J. Vytlacil. 2006. Understanding instrumental variables in models with essential heterogeneity. *Review of Economics and Statistics* 88: 389–432.
Imbens, G.W., and W.K. Newey. 2006. Identification and estimation of triangular simultaneous equations models without additivity. Unpublished manuscript, Department of Economics, MIT.
Manski, C.F. 1988. Identification of binary response models. *Journal of the American Statistical Association* 83: 729–738.
Matzkin, R.L. 1992. Nonparametric and distribution-free estimation of the binary threshold crossing and the binary choice models. *Econometrica* 60: 239–270.
Matzkin, R.L. 2003. Nonparametric estimation of nonadditive random functions. *Econometrica* 71: 1393–1375.
Newey, W.K., J.L. Powell, and F. Vella. 1999. Nonparametric estimation of triangular simultaneous equations models. *Econometrica* 67: 565–603.
Olley, G.S., and A. Pakes. 1996. The dynamics of productivity in the telecommunications equipment industry. *Econometrica* 64: 1263–1297.
Robinson, P.M. 1988. Root-n-consistent semiparametric regression. *Econometrica* 56: 931–954.
Roy, A.D. 1951. Some thoughts on the distribution of earnings. *Oxford Economic Papers* 3: 135–146.
Telser, L.G. 1964. Iterative estimation of a set of linear regression equations. *Journal of the American Statistical Association* 59: 845–862.
Willis, R.J., and S. Rosen. 1979. Education and self-selection. *Journal of Political Economy* 87(5, Par 2): S7–S36.
Wooldridge, J.M. 1997. On two stage least squares estimation of the average treatment effect in a random coefficient model. *Economics Letters* 56: 129–133.
Wooldridge, J.M. 2003. Further results on instrumental variables estimation of average treatment effects in the correlated random coefficient model. *Economics Letters* 79: 185–191.

Conventionalism

Lawrence A. Boland

Abstract

Conventionalism is the methodological doctrine that asserts that explanatory ideas should not be considered true or false but merely better or worse. The truth status of theories cannot be so easily dismissed. While a choice of language may be conventional, the truth status is not a matter of convenient choice. Among economists the most common practice is to avoid using the words 'true' (or 'false') when discussing models and assumptions and instead to invoke 'best' by using a conventionalist theory-choice truth-likeness criterion. The notion of a conventionalist theory-choice criterion presumes a philosophical necessity to choose one theory among competitors.

Keywords

Aumann, R.; Conventionalism; Conventions; Friedman, M.; Hume, D.; Instrumentalism; Lucas, R.; Mathematics and economics; Methodological pluralism; Methodology of economics; Popper, K.; Probability calculus; Problem of induction; Samuelson, P.; Simon, H.; Subjective and objective probability; Testing

JEL Classifications

B4

Conventionalism is the methodological doctrine that asserts that explanatory ideas should not be considered true or false but merely better or worse. At the beginning of the 20th century the status of the laws of physics was the burning issue. It was the famous philosopher Henri Poincaré who in 1902 asked whether the laws of physics were 'only arbitrary conventions'. He answered 'Conventions, yes; arbitrary, no'. Obviously, languages and measurement units are arbitrary conventions but nobody would seriously claim they were explanatory ideas. In Poincaré's day, the question bothering physicists who were dealing with Albert Einstein's new theory (namely, relativity) was whether the choice between Euclidian and non-Euclidian geometry was a matter of convention – that is, a matter of convenience. For everyday questions Euclidian geometry is convenient but perhaps for Einstein's physics non- Euclidian is the better choice. For some matters, such as the choice of language to express

an idea or of units to measure a distance, most people would allow that such a choice may be completely arbitrary.

Although few of them have ever heard of Poincaré, most economists will say almost the same thing whenever they make a methodological pronouncement concerning the truth status of economic theories, models or assumptions. Rarely, however, have economists been concerned with the questions raised about non- Euclidian geometry (except for John Maynard Keynes's metaphorical suggestion at the beginning of his *General Theory*). Of course, hardly any economist questions language being a matter of convenience; moreover, economists often justify the use of mathematics by claiming that its use is like that of language and thus should be judged by its convenience, not its truth status (Samuelson 1952, 1954). But in the 1940s critics of Marshallian and Walrasian (that is, neoclassical) economics argued that the truth status of a theory's assumptions should matter. In his 1953 response to the critics of the realism of assuming perfect competition when explaining the economy, Milton Friedman advocated an alternative methodology: instrumentalism. Instrumentalism, unlike conventionalism, claims merely that the truth status of assumptions does not matter so long as the theory is useful. For those economists who still think the truth status of their theories matters, but realize that one can never prove a theory's truth status by induction, the most common response is something like Poincaré's conventionalism.

There are many examples of economists making methodological pronouncements that exhibit adherence to conventionalism. Paul Samuelson denied that any economic explanation was true, writing that 'An explanation . . . is a better kind of description' (1965, p. 1165). Obviously, some descriptions are better than others, and thus he claimed that we give the honorific title of 'explanation' to the best description. If one were to agree with Samuelson then one certainly could never claim that one's explanation was true. Herbert Simon chose to express this differently; he said all explanations are approximations. Specifically, he said (1963, p. 231) 'Unreality of premises is not a virtue in scientific theory; it is a necessary evil – a concession to the finite computing capacity of the scientist that is made tolerable by the principle of continuity of approximation'. Robert Lucas agreed with that when he said 'Any model that is well enough articulated to give clear answers to the questions we put to it will necessarily be artificial, abstract, patently "unreal"' (1980, p. 696). Robert Aumann, the game theorist, has advocated an even more limited view for explanatory theories. As he put it 'scientific theories are not to be considered "true" or "false"'. Going further, he said, 'In constructing such a theory, we are not trying to get at the truth, or even to approximate to it: rather, we are trying to organize our thoughts and observations in a useful manner.' In this regard, he argued that a theory is like 'a filing system in an office operation, or to some kind of complex computer program' (1985, pp. 31–2). Lucas and Aumann were merely restating Samuelson's 1965 position on methodology.

The Philosophy of Conventionalism

For followers of philosophers Willard Quine and Karl Popper, the truth status of explanations or theories cannot be so easily dismissed or limited. While any choice of language or units of measurement may be conventional, the truth status of theories is not a matter of choice, convenient or otherwise.

Unfortunately, the methodological doctrine of conventionalism is often confused with instrumentalism. As the philosopher Joseph Agassi (1966a) points out, they are responses to two different questions. One concerns the role of theories and the other the truth status of theories. Specifically, if we ask 'What is the *role* of a theory?', instrumentalism's answer is that theories are tools and should not be judged by epistemological standards of truth status or by conventionalist criteria of approximate truth or relative merit (except, perhaps, by simplicity or economy). Conventionalism's different answer is the one stated by Aumann: theories are filing systems or catalogues of observed data. Of course, every

description is also an appeal to a filing system in that one depicts or locates it within a system by referring to other defined dimensions and concepts. If, instead, we chose the question, 'What is the *status* of a theory?', conventionalism's answer is that, of course, theories are approximations and thus should not be considered true or false but better or worse. Instrumentalism's position is simply that truth status does not matter. With this in mind, it is easy to find economists advocating both methodological positions depending on which question is asked. For example, after saying that a theory is like a filing system, Aumann goes on to say that 'We do not refer to such a system as being "true" or "untrue"; rather, we talk about whether it "works" or not, or, better yet, how well it works' (1985, p. 32).

From the perspective of the philosopher Karl Popper, the main question is: what problem is solved by the doctrine of conventionalism? Since the time when Adam Smith's friend David Hume observed that there was no logical justification for the common belief that much of our empirical knowledge was based on inductive proofs (see Russell 1945), methodologists and philosophers have been plagued with what they call the 'problem of induction'. The paradigmatic instance of the problem of induction is the realization that we cannot provide an inductive proof that 'the sun will rise tomorrow'. This leads many of us to ask, 'So *how* do we know the sun will rise tomorrow?' If it is impossible to provide a proof, then presumably we would have to admit we do not know the answer to this burning question! Several writers have claimed to have solved this famous problem (for a discussion of such claims, see Miller 2002). Such a claim is quite surprising since it is a problem that is impossible to solve. Nevertheless, what it is and how it is either 'solved' or circumvented is fundamental to understanding all contemporary methodological discussions.

Up to the time of Popper's entry into the discussion in the mid-1930s, most philosophers took it for granted that all claims to knowledge must be justified. Inductive arguments were seen to be the obvious method. But Popper acknowledged the problem that as a matter of simple logic an inductive argument is impossible. A logical argument is one in which, whenever all the premises are true, any logically derived statements must also be true. An inductive argument is one in which one would argue logically from the truth of particular statements (for example, observation statements such as 'the sun rose today at 7 a.m.') alone to prove the truth of a general statement (for example, the sun always rises). The 'problem of induction' would be solved if one could demonstrate the existence of such an inductive logic. The importance of this problem arises once one realizes that, without some premise of a general nature (such as we find in physics concerning the movement of the earth around the sun and earth's rotation), no finite set of observations could ever prove the non-existence of a counter-example (a refuting instance that would be denied by the general statement in question) somewhere or sometime in the future. For example, to prove that the statement 'All ravens are black' is true requires a proof that there does not exist anywhere in the universe a 'non-black raven'. Everyone agrees that one cannot provide such a literal negative proof. So, it has been argued (Boland 1982, 2003), most discussions of methodology in economics are concerned with the problem *with* induction rather than the problem *of* induction.

Conventionalism can be seen as a solution to the problem *with* induction. Conventionalism presumes that this problem can be solved even though the problem *of* induction cannot. That is, if there were an inductive logic, then the truth status of a true theory or model could in principle be provable since all assumptions of a universal form could be inductively proven. Without such a logic, many think – still insisting that any claim to knowledge must be justified – that some other means must be found to sort through competing theories. That is, how can we choose the best from a set of competing theories? More specifically, by what criteria do we choose between competing theories? Obvious examples of such criteria are simplicity, generality, robustness, testability, falsifiability, verifiability, confirmability, operational meaningfulness, plausibility, probability, and so on. None of these criteria are considered substitutes for truth status (truth or falsity); they are only

choice criteria (truthlikeness). If a criterion can be quantified, one could even see the choice as a matter of applying economics (see Boland 1971). For example, one might choose the theory that is most confirmed – but it still must be remembered that today's most confirmed theory could be a false theory even today.

For many philosophers, such theory-choice criteria are just short-run solutions to the problem *with* induction. That is, in the short run we might be satisfied with invoking such criteria, so that we can choose between theories and thereby be able to push on, but it is hoped that in the long run someone can come up with a solution to the problem *of* induction.

Conventionalism as Employed by Economists

Among economists who openly practise conventionalism, it is a doctrine with many variants and relatives. The most common practice is the avoidance of using the words 'true' (or 'false') when discussing theories, models and assumptions. Instead, we see 'best' being invoked with the use of some conventionalist theory-choice truthlikeness criterion. Also common is the use of the word 'valid' to avoid saying 'true'. Sometimes it is used to mean that a theory is valid if it is logically consistent with available data or evidence. The difficulty is that 'valid' is a question of the logicality of an argument (do the conclusions necessarily follow from the assumptions made?) A logically valid argument can still be false, so it is not always clear what is meant by a valid statement or a valid theory.

One weak form conventionalism is old-fashioned relativism. Another weak form is what the followers of McCloskey (1983) call modernism. In yet another weak form it can be seen to be the rationale for so-called methodological pluralism. The most common form is stronger in that it involves the notion that theories are to be evaluated or compared by means of some form of probability calculus.

Those adherents to conventionalism who advocate the objective form of probability calculus seem unaware of the logical difficulties involved. One might wish to use probability as the measure of confirmation of a theory so that one could use such a measure as the criterion for theory choice. The difficulty arises when one asks what constitutes positive evidence – namely, evidence to be used to calculate the probability measure that would serve as, say, the 'degree of confirmation'. Of course, if one requires all observational evidence to be exactly true, then to be an actual confirmation the objective probability measure would have to be 1.00. That is, just one true observation that contradicts the theory in question requires the rejection of the theory. So it would seem that objective probability measures are inappropriate. But econometrics-based hypothesis testing is not as strict since it allows for errors in the observations of the variables. Hence, the objective probability measure can be of some value less that 1.00. Theory choice in this case would seem to be a simple matter of choosing the theory with the highest probability, that is, the highest degree of confirmation.

Among those who openly advocate a subjective form of probabilities, the most common view is based on Bayesian probabilities which provide a compromise by allowing for explicit roles for both subjectivism in the form of prior probability assessments and objectivism in the form of adjustments based on new objective evidence. Again, the main question for using probabilities concerns what would count as confirming evidence or evidence that increases the subjective probability. Like all confirmation criteria, even if everyone attaches a high subjective probability to the theory in question being true it could still be false and perhaps refuted by the next observation report.

The common element underlying all probability measures to be used for theory choice is the notion that the number of confirming observations should somehow matter. Of course, such an expectation does not require the questionable use of probabilities as a measure of confirmation. But avoiding any reliance on probability will not circumvent the more well-known logical problems of confirmation. All conceptions of a logical connection between positive evidence and degrees of confirmation suffer from a profound logical

problem called, by some philosophers, the 'paradox of confirmation' or the 'paradox of the ravens' (cf. Sainsbury 1995; Agassi 1966b).

The philosopher's paradox of confirmation merely points out that *any* evidence which does not refute a simple universal statement, say, 'All ravens are black' must increase the degree of confirmation. The paradox is based on the observation that, *in terms of what observable evidence would count,* this example of a simple universal statement is logically equivalent to its 'contra-positive' statement 'All non-black things are non-ravens'. Any true observation that is consistent with one of the statements is consistent with the other (equivalent) statement. But in these terms it must be recognized that positive evidence consistent with the contra-positive statement includes red shoes as well as white swans – since in both cases we have non-black things which are not ravens. That is, the set of all confirming instances must include all things which are not non-black ravens. In other words, the more red shoes we observe, the more evidence there is in favour of the contra-positive statement – that is, a red shoe increases the universal statement's degree of confirmation – and, since the contra-positive statement is logically equivalent to the universal statement in question, the latter's degree of confirmation also increases. Obviously, this consideration merely divides the contents of the universe into non-black ravens and everything else (Hempel 1966). This consideration calls into question all claims of confirmation.

Few economists who make pronouncements concerning the appropriate methodology to use in economics are aware of the philosophical problems involved. Almost all think we must have some criterion to choose between competing theories or models. All of them take for granted the necessity of justifying their choice. No recognition seems to be given to the simple fact that one's favourite theory can be true even though it cannot be proven true. That is, whether one's theory is true is a separate question from how one knows it to be true.

The notion of a conventionalist theory-choice criterion presumes that there is a philosophical necessity to choose one theory from among its competitors. But there is no such necessity, even though it will always be difficult to convince economists of this whenever they are naive concerning the philosophy of science. But, given that there are so many different criteria to use, one would think any theory that is best by all criteria should be the chosen theory. But it is doubtful that any theory could satisfy all criteria; so the question is begged as to which criterion is the best criterion. This question seems to put us on the road of an infinite regress: by what criterion do we choose the best criterion to choose between theories? Not a promising journey.

See Also

- ▶ Assumptions Controversy
- ▶ Instrumentalism and Operationalism
- ▶ Pluralism in Economics

Bibliography

Agassi, J. 1966a. Sensationalism. *Mind* 75: 1–24.

Agassi, J. 1966b. The mystery of the ravens: Discussion. *Philosophy of Science* 33: 395–402.

Aumann, R. 1985. What is game theory trying to accomplish? In *Frontiers of economics*, ed. K. Arrow and S. Honkapohja. Oxford: Basil Blackwell.

Boland, L. 1971. Methodology as an exercise in economic analysis. *Philosophy of Science* 38: 105–117.

Boland, L. 1982. *The foundations of economic method*. London: Allen & Unwin.

Boland, L. 2003. *The foundations of economic method: A Popperian perspective*. London: Routledge.

Friedman, M. 1953. Methodology of positive economics. In *Essays in positive economics.* Chicago: University of Chicago Press.

Hempel, C. 1966. *Foundations of natural science*. Englewood Cliffs: Prentice-Hall.

Keynes, J.M. 1936. *General theory of employment, interest and money.* New York: Harcourt, Brace & World.

Lucas, R. 1980. Methods and problems in business cycle theory. *Journal of Money, Credit and Banking* 12: 696–715.

McCloskey, D. 1983. The rhetoric of economics. *Journal of Economic Literature* 21: 481–517.

Miller, D. 2002. Induction: A problem solved. In *Karl Poppers kritischer Rationalismus heute*, ed. J. Böhm, H. Holweg, and C. Hoock. Tübingen: Mohr Siebeck.

Poincaré, H. 1902. *La science et l'hypothèse*. Paris: Flammarion.

Poincaré, H. 1905. *Science and hypothesis*. London: Walter Scott Publishing Company.

Russell, B. 1945. *A history of Western philosophy*. New York: Simon and Schuster.
Sainsbury, R. 1995. *Paradoxes*. Cambridge: Cambridge University Press.
Samuelson, P. 1952. Economic theory and mathematics: An appraisal. *American Economic Review* 42: 56–66.
Samuelson, P. 1954. Some psychological aspects of mathematics and economics. *Review of Economics and Statistics* 36: 380–382.
Samuelson, P. 1965. Professor Samuelson on theory and realism: Reply. *American Economic Review* 55: 1164–1172.
Simon, H. 1963. Problems of methodology: Discussion. *American Economic Review, Proceedings* 53: 229–231.

Convergence

Steven N. Durlauf and Paul A. Johnson

Abstract

One of the most widely studied empirical questions in the new growth economics concerns the role of initial conditions in affecting long-run outcomes. The statistical formulation of this dependence is known as convergence. This article surveys empirical work on convergence, with emphasis on the relationships between conventional definitions of convergence, the main statistical frameworks of evaluating convergence, and various economic models.

Keywords

Cass–Koopmans growth model; Cobb–Douglas functions; Cointegration; Convergence; Endogenous growth; Galton's fallacy; Growth nonlinearities; Identification; Income distribution; Literacy rates; Neoclassical growth theory; Production functions; Solow growth model; Statistics and economics; Technical change; Time series analysis; Regression tree

JEL Classifications

O4

The general question of convergence, understood as the tendency of differences between countries to disappear over time, is of long-standing interest to social scientists. In the 1950s and early 1960s, many analysts discussed whether capitalist and socialist economies would converge over time, in the sense that market institutions would begin to shape socialist economies just as government regulation and a range of social welfare policies grew in capitalist ones.

In modern economic parlance, convergence usually refers specifically to issues related to the persistence or transience of differences in per capita output between economic units, be they countries, regions or states. Most research has focused on convergence across countries, since the large contemporaneous differences between countries generally dwarf intra-country differences. In the context of economic growth, the convergence hypothesis arguably represents the most commonly studied aspect of growth, although the effort to identify growth determinants is arguably the main area of contemporary growth research.

In this overview of convergence, our primary emphasis will be on the development of precise statistical definitions of convergence. This reflects an important virtue of the current literature, namely, the introduction of statistical methods to adjudicate whether convergence is present. At the same time, there is no single definition of convergence in the literature, which is one reason why empirical evidence on convergence is indecisive. Our discussion focuses on convergence across countries, which has dominated empirical studies, although there is reference to studies that focus on other units.

β-Convergence

The primary definition of convergence used in the modern growth literature is based on the relationship between initial income and subsequent growth. The basic idea is that two countries exhibit convergence if the one with lower initial income grows faster than the other. The local (relative to steady state) dynamics of the

neoclassical growth model in both its Solow and Cass–Koopmans variants imply that lower-income economies will grow faster than higher-income ones.

As a statistical question, this notion of convergence can be operationalized in the context of a cross-country regression. Let g_i denote real per capita growth of country i across some fixed time interval and $y_{i,0}$ denote the initial per capita income for country i. Then, unconditional β-convergence is said to hold if, in the regression

$$g_i = k + \log y_{i,0}\beta + \varepsilon_i, \beta < 0. \quad (1)$$

For cross-country regression analysis, one typically does not find unconditional β-convergence unless the sample is restricted to very similar countries, for example, members of the OECD. This finding is in some ways not surprising, since unconditional β-convergence is typically not a prediction of the existing body of growth theories. The reason for this is that growth theories universally imply that growth is determined by factors other than initial income. While different theories may propose different factors, they collectively imply that (1) is misspecified. As a result, most empirical work focuses on conditional β-convergence. Conditional β-convergence holds if $\beta < 0$ for the regression

$$g_i = k + \log y_{i,0}\beta + Z_i\gamma + \varepsilon_i \quad (2)$$

where Z_i is a set of those growth determinants that are assumed to affect growth in addition to a country's initial income. While many differences exist in the choice of controls, it is nearly universal to include those determinants predicted by the Solow growth model, that is, population growth and human and physical capital accumulation rates.

Unlike unconditional β-convergence, evidence of conditional β-convergence has been found in many contexts. For the cross-country case, the basic finding is generally attributed to Barro (1991), Barro and Sala-i-Martin (1992) and Mankiw et al. (1992). The Mankiw, Romer and Weil analysis is of particular interest as it is based on a regression suggested by the dynamics of the Solow growth model. Hence, their findings have been widely interpreted as evidence in favour of decreasing returns to scale in capital (the source of $\beta < 0$ in the Solow model), and therefore as evidence against the Lucas–Romer endogenous growth approach, which emphasizes increasing returns in capital accumulation (either human or physical) as a source of perpetual growth.

From the perspective of the neoclassical growth model, the term $-\beta$ also measures the rate at which an economy's convergence towards its steady-state growth rate, that is, the growth rate determined exclusively by the exogenous rate of technical change. The many findings in the cross-country literature are often summarized by the claim countries converge towards their steady-state growth rates at a rate of about two per cent per year, although individual studies produce different results. The convergence rate has received inadequate attention in the sense that a finding of convergence may have little consequence for questions such as policy interventions if it is sufficiently slow.

As is clear from (2), any claims about conditional convergence necessarily depend on the choice of control variables Z_i. This is a serious concern given the lack of consensus in growth economics on which growth determinants are empirically important. Doppelhofer et al. (2004) and Fernandez et al. (2001) use model averaging methods to show that the cross-country findings that have appeared for conditional β-convergence are robust to the choice of controls. A number of additional statistical issues such as the role of measurement error and endogeneity of regressors are surveyed and evaluated in Durlauf et al. (2005).

The assumption in cross-section growth regressions that the unobserved growth terms ε_i are uncorrelated with log $y_{i,0}$ rules out the possibility that there are country-specific differences in output levels; if such effects were present, they would imply a link between the two. For this reason, a number of researchers have investigated convergence using panel data. This leads to models of the form

$$g_{i,t} = c_i + \log y_{i,t-1}\beta + Z_{i,t}\gamma + \varepsilon_{i,t} \quad (3)$$

where growth is now measured between $t-1$ and t. This approach not only can handle fixed effects, but can allow for instrumental variables to be used

to address endogeneity issues. Panel analyses have been conducted by Caselli et al. (1996), Islam (1995) and Lee et al. (1997). These studies have generally found convergence with rather higher rates than appear in the cross-section studies; for example, Caselli et al. (1996) report annual convergence rate estimates of ten per cent.

As discussed in Durlauf and Quah (1999) and Durlauf et al. (2005), panel data approaches to convergence suffer from the problem that, once country specific effects are allowed, it becomes more difficult to interpret results in terms of the underlying economics. The problem is that, once one allows for fixed effects, then the question of convergence is changed, at least if the goal is to understand whether initial conditions matter; simply put, the country-specific effects are themselves a form of initial conditions. When studies such as Lee et al. (1997) allow for rich forms of parameter heterogeneity across countries, β-convergence become equivalent to the question of whether there is some mean reversion in a country's output process, not whether certain types of contemporaneous inequalities diminish. This does not diminish the interest of these studies as statistical analyses, but means their economic import can be unclear.

σ-Convergence and the Cross-Section Distribution of Income

A second common statistical measure of convergence focuses on the whether or not the cross-section variance of per capita output across countries is or is not shrinking. A reduction in this variance is interpreted as convergence. Letting $\sigma^2_{\log y,t}$ denote the variance across i of $\log y_{it}$, σ-convergence occurs between t and $t+T$ if

$$\sigma^2_{\log y,t} - \sigma^2_{\log y,t+T} > 0. \tag{4}$$

There is no necessary relationship between β- and σ-convergence. For example, if the first difference of output in each country obey $\log y_{i,t} - \log y_{i,t-1} = \beta \log y_{i,t-1} + \varepsilon_{i,t}$, then $\beta < 0$ is compatible with a constant cross-sectional variance (which in this example will equal the variance of $\log y_{i,t}$). The incorrect idea that mean reversion in time series implies that its variance is declining is known as Galton's fallacy; its relevance to understanding the relationship between convergence concepts in the growth literature was identified by Friedman (1992) and Quah (1993a). While it is possible to construct a cross-section regression to test for σ-convergence (cf. Cannon and Duck 2000), they do not test β-convergence per se.

Work on β-convergence has led to general interest in the evolution of the crosscountry income distribution. Quah (1993b, 1996) has been very influential in his modelling of a stochastic process for the distribution itself, with the conclusion that it is converging towards a bimodal steady-state distribution. Other studies of the evolution of the cross-section distribution include Anderson (2004) who uses nonparametric density methods to identify increasing polarization between rich and poor economies across time. Increasing divergence between OECD and non-OECD economies is shown in Maasoumi et al. (2007), working with residuals from linear growth regressions.

One difficulty with convergence approaches that emphasize changes in the shape of the cross-section distribution is that they may fail to address the original question of the persistence of contemporaneous inequality. The reason for this is that it is possible, because of movements in relative position within the distribution, for the cross-section distribution to flatten out while at the same time differences at one point in time are reversed; similarly, the cross-section distribution can become less diffuse while gaps between rich and poor widen. That being said, an examination of the locations of individual countries in various distribution studies typically indicates that the increasing polarization of the world income distribution is mirrored by increasing gaps between rich and poor. A useful extension of this type of research would be to employ the dynamics of individual countries to provide additional information on how the cross-section distribution evolves.

Time Series Approaches to Convergence

An alternative approach to convergence is focused on direct evaluation of the persistence of transitivity of per capita output differences between economies. This approach originates in Bernard and Durlauf (1995), who equate convergence with the statement that

$$\lim_{T\Rightarrow\infty} E\left(\log y_{i,t+T} - \log y_{j,t+T} | F_t\right) = 0 \quad (5)$$

where F_t denotes the history of the two output series up to time t. They find that convergence does not hold for OECD economies, although there is some cointegration in the individual output series. Hobijn and Franses (2000) find similar results for a large international data-set. Evans (1996) employs a clever analysis of the evolution of the cross-section variance to evaluate the presence of a common trend in OECD output, and finds one is present; his analysis allows for different deterministic trends in output and so in this sense is compatible with Bernard and Durlauf (1995).

The relationship between cross-section and time series convergence tests is complicated. Bernard and Durlauf (1996) argue that the two classes of tests are based on different assumptions about the data under study. Cross-section tests assume that countries are in transition to a steady state, so that the data for a given country at time t is drawn from a different stochastic process from the data at some future $t + T$. In contrast, time series tests assume that the underlying stochastic processes are time-invariant parameters, that is, that countries have transited to an invariant output process. They further indicate how convergence under a cross-section test can in fact imply a failure of convergence under a time series test, because of these different assumptions. For these reasons, time series tests of convergence seem appropriate for economies that are at similar stages and advanced stages of development.

From Statistics to Economics

The various concepts of convergence we have described are all purely statistical definitions. The economic questions that motivated these definitions are not, however, equivalent to these questions, so it is important to consider convergence as an economic concept in order to assess what is learned in the statistical studies. As argued in Durlauf et al. (2005), the economic questions that underlie convergence study revolve around the respective roles of initial conditions versus structural heterogeneity in explaining differences in per capita output levels or growth rates. It is the permanent effect of initial conditions, not structural features that matters for convergence. If we define initial conditions as ρ_{i0} and the structural characteristics as $\theta_{i,0}$, convergence can be defined via

$$\lim_{t\to\infty} E\left(\log y_{i,t} - \log y_{j,t} | \rho_{i,0}, \theta_{i,0}, \theta_{j,0}\right) = 0 \text{ if } \theta_{i,0} = \theta_{j,0}. \quad (6)$$

The gap between the definition (6) and the statistical tests that have been employed is evident when one considers whether the statistical tests can differentiate between economically interesting growth models, some of which fulfil (6) and others of which do not. One such contrast is between the Solow growth model and the Azariadis and Drazen (1990) model of threshold externalities, in which countries will converge to one of several possible steady states, with initial conditions determining which one emerges. By definition (6), the Solow model produces convergence whereas the Azariadis–Drazen model does not. However, as shown by Bernard and Durlauf (1996) it is possible for data from the Azariadis–Drazen model to produce estimates that are consistent with a finding of β-convergence.

There is in fact a range of empirical findings of growth nonlinearities that are inconsistent with convergence in the sense of (6). Durlauf and Johnson (1995) is an early study of this type, which explicitly estimated a version of the Azariadis–Drazen model in which the Solow model, under the assumption of a Cobb–Douglas aggregate production function, is a special case. Durlauf and Johnson rejected the Solow model specification and found multiple growth regimes

indexed by initial conditions. Their findings are consistent with the presence of convergence clubs in which different groups of countries are associated with one of several possible steady states. These results are confirmed by Papageorgiou and Masanjala (2004) using a CES production function specification.

The Durlauf and Johnson analysis uses a particular classification procedure, known as a regression tree, to identify groups of countries obeying a common linear model. Other statistical approaches have also identified convergence clubs. For example, Bloom et al. (2003) use mixture distribution methods to model countries as associated with one of two possible output processes, and conclude that individual countries may be classified into high-output manufacturing- and service-based economies and low-output agriculture-based economies. Canova (2004) uses Bayesian methods to identify convergence clubs for European regions.

As discussed in Durlauf and Johnson (1995) and Durlauf et al. (2005), studies of nonlinearity also suffer from identification problems with respect to questions of convergence. One problem is that a given data-set cannot fully uncover the full nature of growth nonlinearities without strong additional assumptions. As a result, it becomes difficult to extrapolate those relationships between predetermined variables and growth to infer steady-state behaviour. Durlauf and Johnson give an example of a data pattern that is compatible with both a single steady and multiple steady states. A second problem concerns the interpretation of the conditioning variables in these exercises. Suppose one finds, as do Durlauf and Johnson, that high- and low-literacy economies are associated with different aggregate production functions. One interpretation of this finding is that the literacy rate proxies for unobserved fixed factors, for example culture, so that these two sets of economies will never obey a common production function, and so will never exhibit convergence in the sense of (6). Alternatively, the aggregate production function could structurally depend on the literacy rate, so that, as literacy increases, the aggregate production functions of currently low-literacy economies will converge to those of the high-literacy ones. Data analyses of the type that have appeared cannot distinguish between these possibilities.

Conclusions

While the empirical convergence literature contains many interesting findings and has helped identify a number of important generalizations about cross-country growth behaviour, it has yet to reach any sort of consensus on the deep economic questions for which the statistical analyses were designed. The fundamentally nonlinear nature of endogenous growth theories renders the conventional crosssection convergence tests inadequate as ways to discriminate between the main classes of theories. Evidence of convergence clubs may simply be evidence of deep nonlinearities in the transitional dynamics towards a unique steady state. Crosssection and time series approaches to convergence not only yield different results but are predicated on different views of the nature of transitory versus steady-state behaviour of economies, differences that themselves have yet to be tested.

None of this is to say that convergence is an empirically meaningless question. Rather, progress requires continued attention to the appropriate statistical definition of convergence and the use of statistical procedures consistent with the definition. Further, it seems important to move beyond current ways of assessing convergence both in terms of better use of economic theory and by a broader view of appropriate data sources. Graham and Temple (2006) illustrate the potential for empirical analyses of convergence that employ well-delineated structural models. The research programme developed in Acemoglu, Johnson and Robinson (2001, 2002) provides a perspective on the micro-foundations of country-specific heterogeneity that speaks directly to the convergence question and which shows the power of empirical analysis based on careful attention to economic history. For these reasons, research on convergence should continue to be productive and important.

See Also

- ▶ Economic Growth, Empirical Regularities In
- ▶ Endogenous Growth Theory
- ▶ Neoclassical Growth Theory
- ▶ Neoclassical Growth Theory (New Perspectives)

Bibliography

Acemoglu, D., S. Johnson, and J. Robinson. 2001. The Colonial origins of comparative development: An empirical investigation. *American Economic Review* 91: 1369–1401.

Acemoglu, D., S. Johnson, and J. Robinson. 2002. Reversal of fortune: Geography and institutions in the making of the modern world income distribution. *Quarterly Journal of Economics* 117: 1231–1294.

Anderson, G. 2004. Making inferences about the polarization, welfare, and poverty of nations: A study of 101 countries 1970–1995. *Journal of Applied Econometrics* 19: 530–550.

Azariadis, C., and A. Drazen. 1990. Threshold externalities in economic development. *Quarterly Journal of Economics* 105: 501–526.

Barro, R. 1991. Economic growth in a cross-section of countries. *Quarterly Journal of Economics* 106: 407–443.

Barro, R., and X. Sala-i-Martin. 1992. Convergence. *Journal of Political Economy* 100: 223–251.

Bernard, A., and S. Durlauf. 1995. Convergence in international output. *Journal of Applied Econometrics* 10(2): 97–108.

Bernard, A., and S. Durlauf. 1996. Interpreting tests of the convergence hypothesis. *Journal of Econometrics* 71(1–2): 161–173.

Bloom, D., D. Canning, and J. Sevilla. 2003. Geography and poverty traps. *Journal of Economic Growth* 8: 355–378.

Canova, F. 2004. Testing for convergence clubs in income per capita: A predictive density approach. *International Economic Review* 45: 49–77.

Cannon, E., and N. Duck. 2000. Galton's fallacy and economic convergence. *Oxford Economic Papers* 53: 415–419.

Caselli, F., G. Esquivel, and F. Lefort. 1996. Reopening the convergence debate: A new look at cross country growth empirics. *Journal of Economic Growth* 1: 363–389.

Doppelhofer, G., R. Miller, and X. Sala-i-Martin. 2004. Determinants of long-term growth: A Bayesian averaging of classical estimates (BACE) approach. *American Economic Review* 94: 813–835.

Durlauf, S., and P. Johnson. 1995. Multiple regimes and cross-country growth behaviour. *Journal of Applied Econometrics* 10: 365–384.

Durlauf, S., P. Johnson, and J. Temple. 2005. Growth econometrics. In *Handbook of economic growth*, ed. P. Aghion and S. Durlauf. Amsterdam: North-Holland.

Durlauf, S., and D. Quah. 1999. The new empirics of economic growth. In *Handbook of macroeconomics*, ed. J. Taylor and M. Woodford. Amsterdam: North-Holland.

Evans, P. 1996. Using cross-country variances to evaluate growth theories. *Journal of Economic Dynamics and Control* 20: 1027–1049.

Fernandez, C., E. Ley, and M. Steel. 2001. Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics* 16: 563–576.

Friedman, M. 1992. Do old fallacies ever die? *Journal of Economic Literature* 30: 2129–2132.

Graham, B., and J. Temple. 2006. Rich nations, poor nations: How much can multiple equilibria explain? *Journal of Economic Growth* 11: 5–41.

Hobijn, B., and P. Franses. 2000. Asymptotically perfect and relative convergence of productivity. *Journal of Applied Econometrics* 15: 59–81.

Islam, N. 1995. Growth empirics: A panel data approach. *Quarterly Journal of Economics* 110: 1127–1170.

Lee, K., M. Pesaran, and R. Smith. 1997. Growth and convergence in multi country empirical stochastic Solow model. *Journal of Applied Econometrics* 12: 357–392.

Maasoumi, E., J. Racine, and T. Stengos. 2007. Growth and convergence: A profile of distribution dynamics and mobility. *Jounal of Econometrics* 136(2): 483–508.

Mankiw, N., D. Romer, and D. Weil. 1992. A contribution to the empirics of economic growth. *Quarterly Journal of Economics* 107: 407–437.

Papageorgiou, C., and W. Masanjala. 2004. The Solow model with CES technology: Nonlinearities with parameter heterogeneity. *Journal of Applied Econometrics* 19: 171–201.

Quah, D. 1993a. Galton's fallacy and tests of the convergence hypothesis. *Scandinavian Journal of Economics* 95: 427–443.

Quah, D. 1993b. Empirical cross-section dynamics in economic growth. *European Economic Review* 37: 426–434.

Quah, D. 1996. Convergence empirics across economies with (some) capital mobility. *Journal of Economic Growth* 1: 95–124.

Convergence Hypothesis

P. J. D. Wiles

This is the doctrine that the Soviet Union and 'similar countries' are becoming and will further become socially and economically similar to the United States and other advanced capitalist countries; or the other way round – so that eventually in

either case political differences, and thus foreign policy tensions, will also disappear.

The doctrine takes many detailed forms, but is most often very unspecific. For instance does it mean: that Texan agriculture will be collectivized (each family farm is larger in area than a Soviet *Kolkhoz*); that there will be a stock exchange again in Moscow, where equity shares in Soviet businesses are freely traded; that the *zloty* will be made convertible; that Switzerland will introduce controls over all retail and wholesale prices; that British trade unions will be reduced to the status of Bulgarian trade unions, or vice versa; that Albania will allow a good deal of minor private enterprise; or even that both sides will converge upon self management in a market, *á la Yougoslave*?

The proponents of the doctrine seldom do it the courtesy of bringing it so close to brass tacks. Above all they fail to recognize just how numerous and diverse those brass tacks are. But the core of the doctrine is clear: advanced capitalism is (said to be) moving, through the large corporation (often public) and its intimacy with certain government departments, irreparably away from share-holder dominance, free enterprise and free markets, in respect of all sectors where small enterprise does not dominate; and a new socio-political type is coming to power, nearly indistinguishable in government and business, and very liable to swap jobs (corruptly, let us add). Meanwhile the advanced Communist states are admitting more and more the role of enterprise independence and markets for everyday small decisions; even the quasi-independence of associations of enterprises in larger decisions – the association would correspond to the corporation and the Communist enterprises to its separate, decentralized 'establishments'.

Hungary and France are of course very much further forward in convergence. A major problem, too for Convergence theorists and for sceptics alike, is China. Here, right at the bottom of the Communist income scale and without even having first introduced any central planning worthy of the name, 20 per cent of the human race is 'converging' very rapidly indeed. As partly too in Hungary, even private property in the means of production is making a comeback. It is not easy to fit this fact into the ordinary framework of debate.

As to a new socio-political type in power in government and 'business', in the USSR ideology is dying and the typical Party apparatchik is more and more obliged to have had some serious professional training and responsibility within the State machine. Meanwhile the obligatory Party membership of the senior technocrat continues to lie lightly on his shoulders. What then is this type, on both sides? It is above all a professional type: technically educated, pragmatic but accepting the particular value system of the given profession, believing in the rule of reason but unphilosophically confusing it with what was judged reasonable at professional school, striving for a higher 'earned' income as the right of competence in his chosen profession, and naive as to what constitutes the rule of reason in unprofessional matters (which are of course the very great majority of matters). One may think in 1986, as the fathers of Convergence certainly did not think, of the American term Yuppie (Young Upwardly Mobile Professionals). However, in the USSR Yuppies are much more idealistic and critical.

It is clear that every prophecy made about Communism in the previous paragraphs is coming true, and the Convergence theorists deserve praise for this – although it took much longer than they expected. The rule of reason is taking over, and the notion that the Soviet system is a frozen monolith, condemned to remain for ever its unpleasant and highly suboptimal (but rapidly growing!) self, is unfounded. But capitalism by no means shows the predicted unilinear change. Japan in one way ('industrial policy', unnaturally accommodating unions) and France in another (mild planning) used to be the showpieces of convergence from the other side. But recent Japanese financial reforms have tended to open up the country to free trade in money, and French planning is at present being down-graded. Monetarist and supply-sider attacks on the public sector and on taxes in the USA and the UK constitute divergence. So does the new tolerance for very heavy unemployment; even if Communist economic experts talk about the necessity for a little

unemployment to discipline labour and create flexibility, the 'target' of Western levels is rapidly receding!

It is, then, capitalism that has 'misbehaved'. And if the rule of reason is eventually restored to economic affairs in the Western world, exactly how far, in so unreasonable a universe, will present divergent trends be reversed? We can at least be sure that protectionism – if that is reason – having flourished even under monetarism, will bloom yet taller under what succeeds it. Indeed under this or that institutional guise, protectionism is common to all systems except capitalist laissez faire in the 19th century. Then too why should not the rule of reason be 'relaxed' again in the East? Besides, 'reason is and ought to be the slave of the passions': if the value systems of Communism and democratic Capitalism continue to diverge only half as much as now, this is cause enough for the reasonable choice of radically divergent policies and substantially divergent institutions.

These considerations alone give us pause before we can accept the basic optimism of Convergence theory. We pause to note that the seven questions of our second paragraph have not been answered at all. But there is worse – though outside of economics – to follow. Since when did resemblance make for peace? Since when was dissimilarity a cause of war? Especially in this ideological age, is not *minor* dissimilarity, or heresy, a major cause of war? For that matter, do not Third World capitalist countries make war on each other, quite unabashedly, over mere boundary disputes and ethnic irredentas in quite the old style? It is very long way from convergence in respect of planning and the market, to international peace.

The alleged *aetiology* of convergence could, as set out above, be the existence of an optimal system somewhere in the middle, to which all existing systems gravitate simply because it is better. If, as is often reasonably claimed, the Yugoslav industrial system represents a third pole of equal theoretical importance, then moderate elements of self-management must be added to that optimal goal. But this is all mere wishful thinking: the judgements of politicians and (where they are counted) voters do not coincide all over the world with each other, let alone with the opinions of centre-left economists. An economic system good for some purposes (e.g. full employment, equality) is bad for others (e.g. rational resource allocation, stable prices, labour discipline). As we have seen, people *value* different sets of outcomes differently, and are also confused as to how in practice to obtain them.

But convergence through contact and competition is another matter. Since nearly all people are unthinking materialists, contact (say as an importer and an exporter) will sway them to imitate the at present more prosperous system: capitalism, to which may or may not be attached, in the perception of observers, parliamentary democracy. And this is truer of people living under Communism than of people in the Third World: for the latter are apt to attribute capitalist prosperity to the exploitation of themselves. Sheer economic contact undoubtedly influences Communist leaders in a capitalist direction, if only because of the overwhelmingly unfavourable balance of technological exchange.

Competition is the almost inevitable result of contact: both commercial and military. It goes without saying that competitors in an export market imitate each other, and not only in quality and technology embodied; but even the administrative systems of the enterprises producing the exports will converge on the one that is seen to be superior. Exporting is a sure and genuine source of convergence, that the most hard-nosed Sovietologist must accept. Military rivalry has much the same effects; for a country's forces also 'export' – a threat. But if the convergence of military technology and its maintenance and auxiliary equipment is of obvious relevance to economic systems, that of military doctrine and organization is not our subject. Still less is the convergence of para-military 'exports': training for guerrillas and terrorists, security systems for underdeveloped countries, espionage.

It can be seen that while high convergence theory is largely (but not altogether) hot air and wishful thinking, there exists a great deal of low-level convergence in fact, all of it easily explicable and much of it very regrettable.

See Also

- ▶ Bureaucracy
- ▶ Command Economy
- ▶ Market Socialism

Convex Programming

Lawrence E. Blume

Abstract

This article summarizes the basic ideas of convex optimization in finite-dimensional vector spaces. Duality, the Fenchel transforms and the subdifferential are introduced and used to discuss Lagrangean duality and the Kuhn–Tucker theorem. Applications of these ideas can be found in duality.

Keywords

Concave optimization; Conjugate duality th; Convex optimization; Convex programming; Convexity; Duality; Fenchel transform; Hyperplanes; Kuhn–Tucker th; Lagrange multipliers; Monotonicity; Quasi-concavity; Saddlepoints; Separation th

JEL Classifications

C68

Introduction

Firms maximize profits and consumers maximize preferences. This is the core of microeconomics, and under conventional assumptions about decreasing returns it is an application of convex programming. The paradigm of convex optimization, however, runs even deeper through economic analysis. The idea that competitive markets perform well, which dates back at least to Adam Smith, has been interpreted since the neoclassical revolution as a variety of conjugate duality for the primal optimization problem of finding Pareto-optimal allocations. The purpose of this article and the companion article duality is (in part) to explain this sentence. This article surveys without proof the basic mathematics of convex sets and convex optimization with an eye towards their application to microeconomic and general equilibrium theory, some of which can be found under duality.

Unfortunately there is no accessible discussion of concave and convex optimization outside textbooks and monographs of convex analysis such as Rockafellar (1970, 1974). Rather than just listing theorems, then, this article attempts to provide a sketch of the main ideas. It is certainly no substitute for the sources. This article covers only convex optimization in finite-dimensional vector spaces. While many of these ideas carry over to infinite-dimensional vector spaces and to important applications in infinite horizon economies and economies with nontrivial uncertainty, the mathematical subtleties of infinite-dimensional topological vector spaces raise issues which cannot reasonably be treated here. The reader looking only for a statement of the Kuhn–Tucker theorem is advised to read backwards from the end, to find the theorem and notation.

A word of warning. This article is written from the perspective of constrained maximization of concave functions because this is the canonical problem in microeconomics. Mathematics texts typically discuss the constrained minimization of convex functions, so textbook treatments will look slightly different.

Convex Sets

A subset C of a Euclidean vector space V is convex if it contains the line segment connecting any two of its members. That is, if x and y are vectors in C and t is a number between 0 and 1, the vector $tx + (1 - t)\,y$ is also in C. A linear combination with non-negative weights which sum to 1 is a *convex combination* of elements of C; a set C is

convex if it contains all convex combinations of its elements.

The key fact about convex sets is the famous *separation th*. A linear function p from the vector space V to $\mathbf{R}$ and a real number a define a *hyperplane*, the solutions to the equation $p \cdot x = a$. Every hyperplane divides V into two *half-spaces*; the upper (closed) half-space, containing those vectors x for which $p \cdot x \geq a$, and the lower (closed) half-space, containing those vectors x for which $p \cdot x \leq a$. The separation theorem uses linear functionals to describe closed convex sets. If a given vector is not in a closed convex set, then there is a hyperplane such that the set lies strictly inside the upper half-space while the vector lies strictly inside the lower half-space:

Separation Theorem If C is a closed convex set and x is not in C, then there is a linear functional p and a real number a such that $p \cdot y > a$ for all $y \in C$, and $p \cdot x < a$.

This theorem implies that every closed convex set is the intersection of the half-spaces containing it. This half-space description is a *dual* description of closed convex sets, since it describes them with linear functionals. From the separation theorem the existence of a *supporting hyperplane* can also be deduced. If x is on the boundary of a closed convex set C, then there is a (non-zero) linear functional p such that $p \cdot y \geq p\, x$ for all $y \in C$; p is the hyperplane that supports C at x.

The origin of the term 'duality' lies in the mathematical construct of the dual to a vector space. The *dual space* of a vector space V is the collection of all *linear functionals*, that is, real-valued linear functions, defined on V. The distinction between vector spaces and their duals is obscured in finite dimensional spaces because each such space is its own dual. If an n-dimensional Euclidean vector space is represented by column vectors of length n, the linear functionals are $1 \times n$ matrices; that is the dual to $\mathbf{R^n}$ is $\mathbf{R^n}$. (This justifies the notation used above.) Self-duality (called reflexivity in the literature) is not generally true in infinite-dimensional spaces, which is reason enough to avoid discussing them here. Nonetheless, although V will be $\mathbf{R^n}$ throughout this article, the usual notation V^* will be used to refer to the dual space of V simply because it is important to know when we are discussing a vector in V and when we are discussing a member of its dual, a linear functional on V.

If the weights in a linear combination sum to 1 but are not constrained to be non-negative, then the linear combination is called an *affine combination*. Just as a convex set is a set which contains all convex combinations of its elements, an affine set in a vector space V is a set which contains all affine combinations of its elements. The set containing all affine combinations of elements in a given set C is an affine set, $A(C)$. The purpose of all this is to define the *relative interior* of a convex set C, ri C. The relative interior of a convex set C is the interior of C relative to $A(C)$. A line segment in $\mathbf{R^2}$ has no interior, but its relative interior is everything on the segment but its endpoints.

Concave Functions

The neoclassical assumptions of producer theory imply that production functions are concave and cost functions are convex. The quasi-concave functions which arise in consumer theory share much in common with concave functions, and quasi-concave programming has a rich duality theory.

In convex programming it is convenient to allow concave functions to take on the value $-\infty$ and convex functions to take on the value $+\infty$. A function f defined on $\mathbf{R^n}$ with range $[-\infty,\infty)$ is concave if the set $\{(x, a) : a \in , a \leq f(x)\}$ is convex. This set, a subset of $\mathbf{R^{n+1}}$, is called the *hypograph* of f and is denoted hypo f. Geometrically, it is the set of points in $\mathbf{R^{n+1}}$ that lie on or below the graph of f. Similarly, the *epigraph* of f is the set of points in $\mathbf{R^{n+1}}$ that lie on or above the graph of f: epi $f = \{(x, a): a \in , a \geq f(x)\}$. A function f with range $(-\infty,\infty]$ is convex $-f$ is concave, and convexity of f is equivalent to convexity of the set epi f. Finally, the *effective domain* of a concave function is the set dom $f = \{x \in \mathbf{R^n} : f(x) > -\infty\}$, and similarly for a convex function. Those familiar with the literature will note that attention here is restricted

to *proper* concave and convex functions. Functions that are everywhere $+\infty$ will also be considered concave, and those everywhere $-\infty$ will be assumed convex when Lagrangeans are discussed below.

Convex optimization does not require that functions be differentiable or even continuous. Our main tool is the separation theorem, and for that closed convex sets are needed. A concave function f is *upper semi-continuous* (usc) if its hypograph is closed; a convex function is *lower semi-continuous* (lsc) if its epigraph is closed. Upper and lower semi-continuity apply to any functions, but these concepts interact nicely conveniently with convex and concave functions. In particular, usc concave and lsc convex functions are continuous on the relative interiors of their domain. A famous example of a usc concave function that fails to be continuous is $f(x, y) = -y^2/2x$ for $x > 0$, 0 at the origin and $-\infty$ otherwise. Along the curve $y = \sqrt{\alpha x},\ y \to 0$ as $x \to 0$, but f is constant at $-\alpha/2$, so f is not continuous at (0,0), but it is usc because the supremum of the limits at the origin is 0.

It is useful to know that, if f is concave and usc, then $f(x) = \inf q(x) = a \cdot x + b$ where the infimum is taken over all a and b such that $a \cdot x + b$ is everywhere at least as big as f. This is another way of saying that, since hypo f is closed, it is the intersection of all half-spaces containing it.

The Fenchel Transform

The concave Fenchel transform associates with each usc function on a Euclidean space V, not necessarily concave, a usc concave function on its dual space V^* (which, we recall, happens to be V since its dimension is finite). The adjective 'concave' is applied because a similar transform is defined slightly differently for convex functions. The concave Fenchel transform of f is

$$f^*(p) = \inf_{x \in V} \{p \cdot x - f(x)\},$$

which is often called the *conjugate* of f. (From here on out we will drop the braces.) The conjugate f^* of f is concave because, for fixed x, $p \cdot x - f(x)$ is linear, hence concave, in p, and the pointwise infimum of concave functions is concave. The textbooks all prove that, if hypo f is closed, so is hypo f^*, that is, upper semi-continuity is preserved by conjugation. So what is this transformation doing, and why is it interesting?

The conjugate f^* of a concave function f describes all the non-vertical half-spaces containing hypo f. This should be checked. A half-space in $\mathbf{R^{n+1}}$ can be represented by the inequality $(p, q)\,(x, y) \geq a$ where q is a real number (as is a) and $p \in V^*$. The half-space is non-vertical if $p^* \neq 0$. In $\mathbf{R^2}$ this means geometrically that the line defining the boundary of the half-space is not vertical. So choose a linear functional $p \neq 0$ in V^*. For any $(x, z) \in$ hypo f, and any $p \in V^*$,

$$p \cdot x - z \geq p \cdot x - f(x) \geq \inf_{x \in V} p \cdot x - f(x) = f^*(p).$$

In other words, the upper half-space $(p, -1) \cdot (x, z) \geq f^*(p)$ contains hypo f. It actually supports hypo f because of the infimum operation: If $a > f^*(p)$, there is an $(x, z) \in$ hypo f such that $px - z < a$, so the upper half-space fails to contain hypo f.

Before seeing what the Fenchel transform is good for, we must answer an obvious qst. If it is good to transform once, why not do it again? Define

$$f^{**}(x) = \inf_{p \in V^*} p \cdot x - f^*(p),$$

the *double dual* of f. The fundamental fact about the Fenchel transform is the following theorem, which is the function version of the dual descriptions of closed convex sets.

Conjugate Duality Theorem If f is usc and concave, then $f^{**} = f$.

This is important enough to explain. Notice that just as p is a linear functional acting on x, so x is a linear functional acting on p. Suppose that f is concave and usc. For all x and p, $p \cdot x - f(x) \geq f^*(p)$,

and so $p \cdot x - f^*(p) \geq f(x)$. Taking the infimum on the left, $f^{**}(x) \geq f(x)$.

On the other hand, take a $p \in V^*$ and a real number b such that the half-space $(p, -1) \cdot (x, z) \geq b$ in $\mathbf{R}^{n+1}$ contains the hypograph of f. This is true if and only if $p \cdot x - b \geq f(x)$ for all x and because f is usc, $f(x)$ is the infimum of $p \cdot x - b$ over all such p and b. Since $p \cdot x - f(x) \geq b$ for all x, take the infimum on the left to conclude that $f^*(p) \geq b$. Thus $p \cdot x - b \geq p \cdot x - f^*(p)$, and taking the infimum now on the right, $p \cdot x - b \geq f^{**}(x)$. Taking the infimum on the left over all the p and b such that the half-space contains hypo f, $f(x) \geq f^{**}(x)$.

It is worthwhile to compute an example to get the feel of the concave Fenchel transform. If C is a closed convex set, the *concave indicator function* of C is $\varphi(x)$ which is 0 for x in C, and $-\infty$ otherwise. This is a good example to see the value of allowing infinite values. The Fenchel transform of φ is $\varphi^*(p) = \inf_{x \in \mathbf{R}^n} p \cdot x - \varphi(x)$. Clearly the infimum cannot be reached at any $x \notin C$, for the value of φ at such an x is $-\infty$, and so the value of $p \cdot x - \varphi(x)$ is $+\infty$. Consequently $\varphi^*(p) = \inf_{x \in C} p \cdot x$. This function has the enticing property of positive homogeneity: If t is a positive scalar, then $\varphi^*(tp) = t\varphi^*(p)$.

Compute the double dual, first for $x \notin C$. The separating hyperplane theorem claims the existence of some p in V^* and a real number a such that $p \cdot x < a \leq p \cdot y$ for all $y \in C$. Take the infimum on the right to conclude that $p \cdot x < \varphi^*(p)$, which is to say $p \cdot x - \varphi^*(p) < 0$. Then, multiply both sides by an arbitrary positive scalar t to conclude that $tp \cdot x - \varphi^*(tp)$ can be made arbitrarily negative. Hence $\varphi^{**}(x) = -\infty$ if $x \notin C$. And if x is in C? Then $p \cdot x - 0 \geq \varphi^*(p)$ for all p (recall $\varphi(x) = 0$). So $p \cdot x - \varphi^*(p) \geq 0$. But $\varphi^*(0) = 0$, so $\varphi^{**}(x)$, the infimum of the left-hand side over all possible p functionals, is 0. Thus the Fenchel transform of φ^* recovers φ.

A particularly interesting version of this problem is to suppose that C is an 'at least as good as' set for level u of some upper semi-continuous and quasi-concave utility function (or, more generally, a convex preference relation with closed weak upper contour sets). Then $\varphi^*(p)$ is just the minimum expenditure necessary to achieve utility u at price p. See duality for more discussion. Another interesting exercise is to apply the Fenchel transform to concave functions which are not usc, and to non-concave functions. These constructions have important applications in optimization theory which we will not pursue.

The theory of convex functions is exactly the same if, rather than the *concave Fenchel transform*, the *convex Fenchel transform* is employed: $\sup_{x \in \mathbf{R}^n} p \cdot x - f(x)$. This transform maps convex lsc functions on V into convex lsc functions on V^*. Both the concave and convex Fenchel transforms will be important in what follows.

The Subdifferential

The separation theorem applied to hypo f implies that usc concave functions have tangent lines: For every $x \in$ ri dom f there is a linear functional p_x such that $f(y) \leq f(x) + p_x(y - x)$. This inequality is called the *subgradient inequality*, and p_x is a *subgradient* of f; p_x defines a tangent line for the graph of f, and the graph lies on or underneath it. The set of subgradients of f at $x \in$ dom f is denoted $\partial f(x)$, and is called the *subdifferential* of f at x. Subdifferentials share many of the derivative's properties. For instance, if $0 \in \partial f(x)$, then x is a global maximum of f. In fact, if $\partial f(x)$ contains only one subgradient p_x, then f is differentiable at x and $Df(x) = p_x$. The set $\partial f(x)$ need not be single-value, however, because f may have kinks. The graph of the function f defined on the real line such that $f(x) = -\infty$ for $x < 0$ and $f(x) = \sqrt{x}$ for $x \geq 0$ illustrates why the subdifferential may be empty at the boundary of the effective domain. At 0, a subgradient would be infinitely steep.

There is a corresponding *subgradient inequality* for convex f: $f(y) \geq f(x) + p_x \cdot (y - x)$. With these definitions, $\partial(-f)(x) = -\partial f(x)$. Note that some texts refer to superdifferentials for concave functions and subdifferentials for convex functions. Others do not multiply the required terminology, and we follow them.

The multivalued map $x \mapsto \partial f(x)$, is called the *subdifferential correspondence* of f. An important property of subdifferential correspondences is

monotonicity. From the subgradient inequality, if $p \in \partial f(x)$ and $q \in \partial f(y)$, then $f(y) \leq f(x) + p \cdot (y - x)$ and $f(x) \leq f(y) + q \cdot (x - y)$, and it follows that $(p - q) \cdot (x - y) \leq 0$. For convex f the inequality is reversed.

The Fenchel transforms establish a clear relationship between the subdifferential correspondences of concave functions and their duals. If f is concave, then the subdifferential inequality says that $p \in \partial f(x)$ if and only if for all $z \in X$, $p \cdot x - f(x) \leq p \cdot z - f(z)$. The map $z \mapsto p \cdot z - f(z)$ is minimized at $z = x$, and so p is in $\partial f(x)$ if and only if $f^*(p) = p \cdot x - f(x)$. If f is usc, then $f^{**} \equiv f$, and so $f^{**}(x) = f(x) = p \cdot x - f^*(x)$. That is, $p \in \partial f(x)$ if and only if $x \in \partial f^*(p)$.

Optimization and Duality

Economics most often presents us with constrained maximization problems. Within the class of problems with concave objective functions, there is no formal difference between constrained and unconstrained maximization. The constrained problem of maximizing concave and usc f on a closed convex set C is the same as the unconstrained problem of maximizing $f(x) + I_C(x)$ on $\mathbf{R^n}$, where $I_C(x)$ is the concave indicator function of C.

The general idea of duality schemes in optimization theory is to represent maximization (or minimization) problems as half of a minimax problem which has a saddle value. There are several reasons why such a seemingly odd construction can be useful. In economics it often turns out that the other half of the minimax problem, the dual problem, sheds additional light on properties and inpts of the primal problem. This is the source of the 'shadow price' concept: The shadow price is the value of relaxing a constraint. Perhaps the most famous example of this is the Second Theorem of Welfare Economics.

Lagrangeans

The *primal problem* (problem P) is to maximize $f(x)$ on a Euclidean space V. Suppose there is a function $L : V \times V^* \to \mathbf{R}$ such that $f(x) = \inf_{p \in V^*} L(x, P)$. Define $g(p) = \sup_{x \in V} L(x, p)$, and consider the problems of maximizing $f(x)$ on V and minimizing $g(p)$ on V^*. The first problem is the primal problem, and the second is called the *dual problem*. For all x and p it is clear that $f(x) \leq L(x, p) \leq g(p)$, and thus that

$$\sup_x \inf_p L(x,p) = \sup_x f(x) \leq \inf_p g(p) = \inf_p \sup_x L(x,p).$$

If the inequality is tight, that is, it holds with equality, then the common value is called a *saddle value* of L. In particular, a saddle value exists if there is a *saddle point* of L, a pair (x^*, p^*) such that for all $x \in V$ and $p \in V^*$, $L(x, p^*) \leq L(x^*, p^*) \leq L(x^*, p)$. A pair (x^*, p^*) is a saddlepoint if and only if x^* solves the primal, p^* solves the dual, and a saddle value exists. The function L is the *Lagrangean*, which is familiar from the analysis of smooth constrained optimization problems. Here it receives a different foundation.

The art of duality schemes is to identify an interesting L, and here is where the Fenchel transforms come in. Interesting Lagrangeans can be generated by embedding the problem max f in a parametric class of concave maximization problems. Suppose that there is a (Euclidean) parameter space P, and a usc and concave function F: $V \times Y \to$ such that $f(x) = F(x, 0)$, and consider all the problems $\max_{x \in V} F(x, y)$. A particularly interesting object of study is the *value function* $\varphi(y) = \sup_x F(x, y)$, which is the indirect utility function in consumer theory, and the cost function in the theory of the firm (with concave replaced by convex and max by min). The map $y \mapsto -F(x, y)$ is closed and convex for each x, so define on $V \times V^*$

$$L(x,p) = \sup_y p \cdot y + F(x,y),$$

its (convex) Fenchel transform. The map $p \mapsto L(x, p)$ is closed and convex on V^*. Transform again to see that $F(x, y) = \inf_{p \in V^*} L(x, p) - p \cdot y$. In particular, $f(x) = \inf_p L(x, p)$.

An example of this scheme is provided by the usual concave optimization problem given by a

concave objective function f, K concave constraints $g_k(x) \geq 0$, and an implicit constraint $x \in C$: $\max_x f(x)$ subject to the constraints $g_k(x) \geq 0$ for all k and $x \in C$. Introduce parameters y so that $g_k(x) \geq y_k$, and define $F(x, y)$ to be $f(x)$ if all the constraints are satisfied and $-\infty$ otherwise. The supremum defining the Lagrangean cannot be realized for y such that x is infeasible, and so

$$L(x,p) = \begin{cases} f(x) + \sum_k p_k g_k(x) & \text{if } x \in C \text{ and } p \in \mathbf{R}^{\mathbf{K}}_+, \\ +\infty & x \in C \text{ and } p \notin \mathbf{R}^{\mathbf{K}}_+ \\ -\infty & \text{if } x \notin C \end{cases} \tag{1}$$

if there are feasible x, then $F(x,y)$ is everywhere $-\infty$, and so $L(x, p) \equiv -\infty$.

Here, in summary, are the properties of the Lagrangean for the problems discussed here:

Lagrangean Theorem If $F(x, y)$ is lsc and concave then (1) the Lagrangean L is lsc and convex in p for each $x \in V$, (2) L is concave in x for each $p \in V^*$, and (3) $f(x) = \inf_p L(x, p)$.

Following the original scheme, the objective for the dual problem is $g(p) = \sup_x L(x, p)$, and the *dual problem* (problem D) is to maximize g on V^*. Perhaps the central fact of this dual scheme is the relationship between the dual objective function g and the value function φ. The function φ is easily seen to be concave, and simply by writing out the definitions, one sees that $g(p) = \sup_y \cdot p \cdot y + \varphi(y)$, the convex Fenchel transform of the convex function $-\varphi$. So $g(p)$ is lsc and convex, $g(p) = (-\varphi)^*(p)$ and whenever φ is usc, $\inf_p g(p) = \varphi(0)$.

To make the duality scheme complete, the min problem should be embedded in a parametric class of problems in a complementary way. Take $G(p, q) = \sup_{x \in V} L(x, p) - q \cdot x$, so that $g(p) = G(p, 0)$. With this definition, $-G(p, q) = \inf_{x \in V} q \cdot x - L(x, p)$, the concave Fenchel transform of $x \mapsto L(x, p)$. The value function for the parametric class of minimization problems is $\gamma(q) = \inf_p G(p, q)$. The relationship between F and G is computed by combining the definitions:

$$\begin{aligned} \mathrm{G}(p,\mathrm{q}) &= \sup_{x,y} F(x,y) - q \cdot x + p \cdot y \\ &= -\inf_{x,y} q \cdot x - p \cdot y - F(x,y) \\ &= -F^*(q, -y) \end{aligned}$$

and so

$$F(x,y) = \inf_{x,y} G(p,q) + q \cdot x - p \cdot y$$

where the F^* is the concave Fenchel transform of the map $(x, y) \mapsto F(x, y)$. Computing from the definitions, $f(x) = \inf_{p,q} q \cdot x + G(p, q) = \inf_q q \cdot x + \gamma(q)$, so $f = (-\gamma)^*$, and whenever γ is lsc, $\sup_x f(x) = \gamma(0)$.

In summary, if $F(x, y)$ is concave in its arguments, and usc, then we have constructed a Lagrangean and a dual problem of minimizing a concave and lsc $G(p, q)$ over p. If the value functions $\varphi(y)$ and $\gamma(q)$ are usc and lsc, respectively, then $\sup_x F(x, 0) = \gamma(0)$ and $\inf_p G(p, 0) = \varphi(0)$, so a saddle value exists. Upper and lower semi-continuity of the value functions can be an issue. The hypograph of φ is the set of all pairs (y, a) such that $\sup_x F(x, y) \geq a$, and this is the projection onto y and a of the set of all triples (x, y, a) such that $F(x, y) \geq a$, that is, hypo F. Unfortunately, even if hypo F is closed, its projection may not be, so upper semi-continuity of φ does not follow from the upper semi-continuity of F.

In the constrained optimization problem with Lagrangean (1), the parametric class of dual minimization problems is to minimize $G(p, q) = \sup_{x \in C} f(x) + \sum_k p_k g_k(x) - q \cdot x$ if $y \in \mathbf{R}^{\mathbf{K}}_+$ and $+\infty$ otherwise. Specialize this still further by considering linear programming. The canonical linear program is to *max* $a \cdot x$ subject to the explicit constraints $b_k \cdot x \leq c_k$ and the implicit constraint $x \geq 0$. Rewrite the constraints as $-b_k \cdot x + c_k \geq 0$ to be consistent with the formulation of (1). Then

$$\begin{aligned} G(p,q) &= \sup_{x \geq 0} \; a \cdot x - \sum_k p_k (\mathrm{b}_k \cdot x - c_k) - q \cdot x \\ &= \sum_k c_k p_k + \sup_{x \geq 0} \left(a - \sum_k p_k b_k - q \right) \cdot x \end{aligned}$$

for $p \in \mathbf{R}^{\mathbf{K}}_{+}$ and $+\infty$ otherwise. The dual problem is to minimize this over p. The sup term in G will be $+\infty$ unless the vector in parentheses is non-positive, in which case the sup will be 0. So the dual problem, taking $q = 0$, is to minimize $\sum_k c_k p_k$ over p subject to the constraints that $\sum_k p_k b_k \geq a$ and $p \in \mathbf{R}^{\mathbf{K}}_{+}$. If the prima constraints are infeasible, $\varphi(0) = -\infty$. If the dual is infeasible, $\gamma(0) = +\infty$, and this serves as an example of how the dual scheme can fail over lack of continuity. For linear programs there is no problem with the hypographs of φ and γ, because these are *polyhedral convex sets*, the intersection of a finite number of closed half-spaces, and projections of closed polyhedral convex sets are closed.

Solutions

Subdifferentials act like partial derivatives, particularly with respect to identifying maxima and minima: x^* in V solves problem P if and only if $0 \in \partial f(x^*)$. When f is identically $-\infty$, there are no solutions which satisfy the constraints. Thus dom f is the set of *feasible solutions* to the primal problem P. Similarly, $p^* \in V^*$ solves the dual problem D if and only if $\partial g(p^*) = 0$, and here dom g is the set of *dual-feasible solutions*. Saddlepoints of the Lagrangean also have a subdifferential characterization. Adapting the obvious partial differential notion and notation, (x^*, p^*) is a saddle point for L if and only if $0 \in \partial_x L(x^*, p^*)$ and $0 \in \partial_p L(x^*, p^*)$ (these are different 0's since they live in different spaces), which we write $(0, 0) \in \partial L(x^*, p^*)$. This condition is often called the *Kuhn–Tucker* condition. The discussion so far can be summarized in the following theorem, which is less general than can be found in the sources:

Kuhn–Tucker theorem Suppose that $F(x, y)$ is concave and usc. Then the following are equivalent:

1. $\sup f = \inf g$,
2. φ is usc and concave,
3. the saddle value of the Lagrangean L exists,
4. γ is lsc and convex.

In addition, the following are equivalent:

5. x^* solves P, p^* solves D, and the saddle value of the Lagrangean exists.
6. (x^*, p^*) satisfy the Kuhn–Tucker condition.

For economists, the most interesting feature of the dual is that it often describes how the value of the primal problem will vary with parameters. This follows from properties of the subdifferential and the key relation between the primal value function and the dual objective function, $g = (-\varphi)^* : -\partial\varphi(0) = \partial(-\varphi)(0)$, and this equals the set $\{p : (-\varphi)^*(p) = p \cdot 0 - (-\varphi)(0)\}$, and this is precisely the set $\{p : g(p) = \sup_x f(x)\}$. In words, if p is a solution to the dual problem D, then $-p$ is in the subgradient of the primal value function. When $\partial\varphi(0)$ is a singleton, there is a unique solution to the dual, and it is the derivative of the value function with respect to the parameters. More generally, from the subdifferential of a convex function one can construct directional derivatives for particular changes in parameter values. Similarly, $-\partial\gamma(0) = \{x : f(x) = \inf_p g(p)\}$, with an identical inpt. In summary, add to the Kuhn–Tucker theorem the following equivalence:

7. $-p^* \in \partial\varphi(0)$ and $-x^* \in \partial\gamma(0)$

The remaining question is, when is any one of these conditions satisfied? A condition guaranteeing that the subdifferentials are non-empty is that $0 \in$ ri dom φ, since concave functions always have subdifferentials on the relative interior of their effective domain. In the constrained optimization problem whose Lagrangean is described in (1), an old condition guaranteeing the existence of saddlepoints is the *Slater condition*, that there is an $x \in$ ri C such that for all k, $g_k(x) > 0$. This condition implies that $0 \in$ ri dom φ, because there is an open neighbourhood around 0 such that for y in the neighbourhood and for all k, $g_k(x) > y_k$. Thus $\varphi(y) \geq F(x, y) > -\infty$ for all y in the neighbourhood. Conditions like this are called *constraint qualifications*. In the standard calculus approach to constrained optimization, they give conditions under which derivatives sufficiently characterize the constraint set for calculus approximations to work (see Arrow et al. 1961).

Finally, it is worth noting that infinite dimensional constrained optimization problems, such as those arising in dynamic economic models and the study of uncertainty, can be addressed with extensions of the methods discussed here. The main difficulty is that most infinite dimensional vector spaces are not like $\mathbf{R^n}$. There is no 'natural' vector space topology, and which topology one chooses has implications for demonstrating the existence of optima. The existence of separating hyperplanes is also a difficulty in infinite dimensional spaces. These and other problems are discussed in Mas-Colell and Zame (1991). Nonetheless, much of the preceding development does go through. See Rockafellar (1974).

See Also

- ▶ Convexity
- ▶ Duality
- ▶ Lagrange Multipliers
- ▶ Quasi-concavity

Bibliography

Arrow, K., L. Hurwicz, and H. Uzawa. 1961. Constraint qualifications in maximization problems. *Naval Logistics Research Quarterly* 8: 175–191.

Mas-Colell, A., and W. Zame. 1991. Equilibrium theory in infinite dimensional spaces. In *Handbook of mathematical economics*, ed. W. Hildenbrand and H. Sonnenschein, vol. 4. Amsterdam: North-Holland.

Rockafellar, R.T. 1970. *Convex analysis*. Princeton: Princeton University Press.

Rockafellar, R.T. 1974. *Conjugate duality and opttimization*, CBMS Regional Conference Series No. 16. Philadelphia: SIAM.

Convexity

Lawrence E. Blume

Keywords

Convexity; Duality; Existence of equilibrium; Hyperplanes; Large economies; Lyapunov's theorem; Optimization; Quasi-concavity; Quasi-convexity; Shapley–Folkman theorem

JEL Classifications

D0

Convexity is the modern expression of the classical law of diminishing returns, which was prominent in political economy from Malthus and Ricardo through the neoclassical revolution. Its importance today rests less on any utilitarian or behavioural psychological rationale or physical principle than on its utility as a tool of mathematical analysis. In general equilibrium and game theory, proofs of the existence of equilibrium, competitive and Nash, respectively, rely on the application of a fixed-point theorem to a set-valued, convex-valued map from a convex set to itself. Welfare economics provides another example: The second theorem of welfare economics, which asserts that optimal allocations can be supported by competitive prices, relies on an application of the supporting hyperplane theorem to an appropriate convex set.

Convexity is a property of real vector spaces, and its domain of application in economic analysis is not just Euclidean spaces but also the infinite dimensional vector spaces which arise in the study of uncertainty and dynamics, where infinite numbers of goods are required. Nonetheless, this brief exposition will be confined to Euclidean spaces.

Definitions

A set $C \subset \mathbf{R^n}$ is *convex* if the line segment connecting any two points in C lies wholly within C. Formally put, C is convex if and only if for all points x and y in C and all scalars t in the unit interval $[0, 1]$, the point $tx + (1 - t)y$ is also in C. A ball is convex; a boomerang is not. An extended real-valued function f defined on a convex set $C \subset \mathbf{R^n}$ is *convex* if its *epigraph* or *supergraph*, $\{(x, \mu) : x \in C, \mu \in \mathbf{R}, f(x) \leq \mu\}$ is convex. For real-valued functions, this is equivalent to the more familiar definition that for all x and y in C and $t \in [0, 1]$, $f(tx + (1 - t)y) \geq tf(x) + (1 - t)f(y)$. A function f is *concave* if $-f$ is convex.

Optimization

Students of economics first encounter convexity in the study of optimization. If $x^* \in \mathbf{R^n}$ is a critical point of a smooth function, and if x^* a local maximum, then the Hessian matrix at x^*, the matrix of second-order partial derivatives, must be negative semi-definite; that is, it is locally concave. Any critical point with a negative definite Hessian must be a local maximum. Negative definiteness of the Hessian implies but is not implied by strict (local) concavity. For Jevons, utility was additively separable, and so the principle of diminishing marginal utility itself was enough to derive concavity. Edgeworth, the first economist to consider non-separable utility functions, realized that diminishing marginal utility was not, in general, enough to guarantee convexity. His development of demand theory relied on a differential condition that can be shown to imply *quasi-concavity.* A real-valued function f with a convex domain C is *quasi-concave* if for each real number α, the set $\{x \in C : f(x) \geq \alpha\}$ is convex. To appreciate the difference between concavity and quasi-concavity, note that any strictly increasing function on the real line is quasi-concave. The differential description of convexity and its variants (quasi-convexity, pseudo-convexity) and the associated necessary and sufficient second-order conditions for constrained optimization problems has produced a volume of analysis, most of which is of second-order importance to contemporary economic theory. Exhaustive coverage can be found in Simon and Blume (1994).

Duality

The representation of consumers by expenditure functions and firms by profit functions is said to be 'dual' to the 'primal' representations by preferences and production sets, respectively. These representations rely on alternative ways of representing closed convex sets: The 'primal' description is a list of its elements, and the 'dual' description is the list of closed half-spaces containing it. The dual representation for closed convex sets is equivalent to the separating hyperplane theorem: If x in $\mathbf{R^n}$ is not in a closed convex set C, then there is a hyperplane $H \subset \mathbf{R^n}$ with x on one side and C on the other. That is, there is a $p \in \mathbf{R^n}$ and a number α such that $p \cdot x < \alpha$ and $p \cdot y > \alpha$ for all $y \in C$. (See ▶ Convex Programming and ▶ Duality.)

Large Numbers and Convexity

Convexity is sometimes an inappropriate assumption. Half a box of two left shoes and half a box of two right shoes is surely preferred to either box, but the 50:50 mixture of a good burgundy and a good stout is only a headache. Fortunately, the analysis of perfectly competitive markets rests not on the preferences of any individual consumer, but on the average behaviour of a large number of consumers. A central insight behind much research of the 1970s and 1980s (and which was anticipated by Edgeworth 1881, a century before) is that averaging is a convexifying operation. This is the content of the Shapley–Folkman theorem as applied to large finite economies, and Lyapunov's theorem in the analysis of economies with a continuum of agents. (See ▶ Cores, ▶ Large Economies and ▶ Perfect Competition.) For economies with large numbers of small consumers and small firms, the important analytical constructs are approximately convex. With respect to the existence of equilibrium and its welfare properties, large economies look like convex economies. Hildenbrand (1974) is an entry point to this important body of research.

See Also

- ▶ Convex Programming
- ▶ Cores
- ▶ Duality
- ▶ Large Economies
- ▶ Perfect Competition

Bibliography

Edgeworth, F.Y. 1881. *Mathematical psychics*. London: C. Kegan Paul & Co.

Hildenbrand, W. 1974. *Core and equilibria of a large economy*. Princeton: Princeton University Press.
Simon, C., and L. Blume. 1994. *Mathematics for economists*. New York: W.W. Norton.

Convict Labour

Farley Grubb

Keywords
Colonialism; Convict labour; International migration; Labour contracts; Migration, international

JEL Classification
N30; N31

Some European countries banished convicts to labour in overseas colonies – sometimes using private markets to transport and employ this labour.

Punishing felons who did not warrant execution and were too poor to pay monetary fines posed a dilemma for early modern societies. The long-standing punishments of one-off physical chastisements, such as whippings, increasingly seemed too barbaric and returned malefactors to society too quickly. While long-term incarceration was more civilized and removed malefactors from society, penitentiaries were expensive to build and operate, and the criminal's labour was lost to society. Sentencing felons to labour in overseas colonies thus became an attractive solution.

Between 1854 and 1920 France sent between 20,000 and 30,000 convicts to French Guiana and New Caledonia. Spain sent convicts to North Africa, Cuba, and Puerto Rico. Britain, however, was the largest participant, sending 6000–10,000 convicts to its colonies between 1614 and 1718 and another 50,000 mostly to its American colonies Virginia and Maryland between 1718 and 1775 (Coldham 1992; Ekirch 1987). After the United States closed its shores to British convicts, convict transportation was shifted to Australia where approximately 160,000 were landed between 1787 and 1868 (Nicholas 1988). Another 18,000 were shipped to Bermuda and Gibraltar.

The Transportation Act of 1718 shifted the overseas banishment of British felons from a case-by-case petitioning of the Crown to a routine sentence imposed by courts. The sentences allowed were 7 years, 14 years, or a lifetime of banishment – 74%, 24% and 2%, respectively, of those transported – which became the length of the convict's overseas labour contract. Most transported convicts were guilty of property crimes and were Englishmen. Between 13 and 23% were Irish, and between 10 and 15% were female (Ekirch 1987). Sentences were not rigidly tied to crimes; for example, highway robbers received 7-year, 14-year, and lifetime sentences – 38%, 50%, and 12%, respectively (Grubb 2000). Not until convicts had completed their sentences could they return to Britain without facing being hanged if caught.

The privatization of overseas convict disposal reached its zenith after the Transportation Act. The government minimized its cost of overseas convict disposal by channelling convicts through the existing competitive markets for voluntary servant labour, where emigrants traded forward-labour contracts to shippers for passage to America. Shippers recouped their cost by selling these contracts (emigrants) to private employers in America. Potential shipping profits related to labour heterogeneity were arbitraged away by bargaining over contract length. The typical voluntary servant negotiated a four-year labour contract and sold in America for eight and half pounds sterling. By contrast, courts fixed the length of convict sentences (labour contracts) independently of labour heterogeneity. Convicts were then transferred to private shippers for transportation overseas. Shippers sold their convicts as servant labour to private employers in America to recoup their shipping expense. The average convict sold for 11 lb sterling (Grubb 2000).

By fixing contract lengths – the parameter used to arbitraged shipping profits in the voluntary servant market – the courts altered the convict auction price distribution and profit arbitrage process from that which existed in the voluntary servant market. The distribution of convict contract prices had a higher mean, higher standard deviation, and lower kurtosis than that of voluntary servant contract prices. Shippers did not earn excess profits on convicts. The higher sale price was matched by the higher cost of chaining convicts during shipment and paying variable fees charged by county jailers. Jailers played shippers off against each other for access to convict cargo. The government subsidized one shipper in the London market who earned, net of political bribes, excess profits (Grubb 2000).

Shippers carried both voluntary and convict servants concurrently. Potential employers were shown the conviction papers that stated each convict's sentence and crime. Post-auction convicts were largely indistinguishable from voluntary servants. Most were employed in agriculture and at iron forges alongside slaves and voluntary servants. They lived in their employer's house and ate at their employer's table. Criminal conviction, however, carried a stigma that led to price discounts. A year's worth of convict labour sold for a 21% discount on average over that of comparable voluntary servant labour. Convicts guilty of more serious and professional crimes, such as arsonists and receivers of stolen goods, sold for even greater discounts. Convicts also ran away more often than did voluntary servants: 16% versus 6%, respectively (Grubb 2001).

Per given crime, a 14-year versus a 7-year sentence signalled the courts' perception of the severity of harm inflicted by, and incorrigibility of, the convict. American employers responded to this information by demanding additional price discounts of 48% and 68% per year of labour for convicts sentenced to 14 years and to life, respectively, as opposed to seven years for the same crime. Employers also paid premiums and received discounts for certain convict attributes, other things equal. For example, taller convicts sold for a substantial premium, and female convicts with venereal disease sold for 19% less than females without the disease (Grubb 2001).

For underpopulated colonies lacking competitive labour markets, such as Australia, European governments typically had to transport convicts to the colonies themselves, directly employing them on government projects there (Nicholas 1988). During the nineteenth century, European governments also became increasingly reluctant to use existing competitive markets to auction convict labour for fear that it would look like government-sanctioned slavery. Instead, convicts were transferred via bureaucratic petition or assignment systems. Under these conditions, the system struggled to employ convict labour efficiently and to be a cost-effective punishment. Convict transportation waned as social reformers succeeded in replacing it with incarceration in newly built penitentiaries and as maturing colonies increasingly resisted being convict dumping-grounds.

See Also

- Auctions (Empirics)
- Compensating Differentials
- Human Capital, Fertility and Growth
- Indentured Servitude
- International Migration
- Labour Market Institutions

Bibliography

Coldham, P. 1992. *Emigrants in chains*. Baltimore: Genealogical Publishing.

Ekirch, A. 1987. *Bound for America: The transportation of British convicts to the colonies, 1718–1775*. New York: Oxford University Press.

Grubb, F. 2000. The transatlantic market for British convict labor. *Journal of Economic History* 60: 94–122.

Grubb, F. 2001. The market evaluation of criminality: Evidence from the auction of British convict labor in America, 1767–1775. *American Economic Review* 91: 295–304.

Nicholas, S. (ed.). 1988. *Convict workers: Reinterpreting Australia's past*. New York: Cambridge University Press.

Cooperation

Samuel Bowles and Herbert Gintis

Abstract

We review game-theoretic models of cooperation with self-regarding agents. We then study the folk theorem in large groups of self-regarding individuals with imperfect information. In contrast to the dyadic case with perfect information, the level of cooperation deteriorates with larger group size and higher error rates. Moreover, no plausible account exists of how the dynamic, out-of-equilibrium behaviour of these models would support cooperative outcomes. We then analyse cooperation with other-regarding preferences, finding that a high level of cooperation can be attained in large groups and with modest informational requirements, and that conditions allowing the evolution of such social preferences are plausible.

Keywords

Cooperation; Focal rules; Folk theorem; Game theory; General equilibrium; Multiple equilibria; Prisoner's Dilemma; Private information; Public goods game; Reciprocity, indirect; Reciprocity, strong; Repeated games; Reputation maintenance; Retaliation; Social preferences; Strategic interaction; Subgame perfection; Tit for tat

JEL Classifications

C9

Cooperation is said to occur when two or more individuals engage in joint actions that result in mutual benefits. Examples include the mutually beneficial exchange of goods, the payment of taxes to finance public goods, team production, common pool resource management, collusion among firms, voting for income redistribution to others, participating in collective actions such as demonstrations, and adhering to socially beneficial norms.

A major goal of economic theory has been to explain how wide-scale cooperation among self-regarding individuals occurs in a decentralized setting. The first thrust of this endeavour involved Walras's general equilibrium model, culminating in the celebrated 'invisible hand' theorem of Arrow and Debreu (Arrow and Debreu 1954; Debreu 1959; Arrow and Hahn 1971). But, the assumption that contracts could completely specify all relevant aspects of all exchanges and could be enforced at zero cost to the exchanging parties is not applicable to many important forms of cooperation. Indeed, such economic institutions as firms, financial institutions, and state agencies depend on incentive mechanisms involving strategic interaction in addition to explicit contracts (Blau 1964; Gintis 1976; Stiglitz 1987; Tirole 1988; Laffont 2000).

The second major thrust in explaining cooperation eschewed complete contracting and developed sophisticated repeated game-theoretic models of strategic interaction. These models are based on the insights of Shubik (1959), Taylor (1976), Axelrod and Hamilton (1981) and others that repetition of social interactions plus retaliation against defectors by withdrawal of cooperation may enforce cooperation among self-regarding individuals. A statement of this line of thinking, applied towards understanding the broad historical and anthropological sweep of human experience is the work of Ken Binmore (1993, 1998, 2005). For Binmore, a society's moral rules are instructions for behaviour in conformity with one of the myriad of Nash equilibria of a repeated n-player social interaction. Because the interactions are repeated, and these rules form a Nash equilibrium, the self-regarding individuals who comprise the social order will conform to the moral rules.

We begin by reviewing models of repeated dyadic interaction in which cooperation may occur among players who initially cooperate and in the next round adopt the action of the other player in the previous round, called *tit for tat*. These models show that as long as the probability of game repetition is sufficiently great and

individuals are sufficiently patient, a cooperative equilibrium can be sustained once it is implemented. This reasoning applies to a wide range of similar strategies. We then analyse *reputation maintenance* models of dyadic interaction, which are relevant when individuals interact with many different individuals, and hence the number of periods before a repeat encounter with any given individual may be too great to support the tit-for-tat strategy.

We then turn to models of cooperation in larger groups, arguably the most relevant case, given the scale on which cooperation frequently takes place. The folk theorem (Fudenberg and Maskin 1986) shows that, in groups of any size, cooperation can be maintained on the assumption that the players are sufficiently future-oriented and termination of the interaction is sufficiently unlikely. We will see, however, that these models do not successfully extend the intuitions of the dyadic models to many-person interactions. The reason is that the level of cooperation that may be supported in this way deteriorates as group size increases and the probability of either behavioural or perceptual error rises, and because the theory lacks a plausible account of how individuals would discover and coordinate on the complicated strategies necessary for cooperation to be sustained in these models. This difficulty bids us investigate how other-regarding preferences, *strong reciprocity* in particular, may sustain a high level of cooperation, even with substantial errors and in large groups.

Repetition Allows Cooperation in Groups of Size Two

Consider a pair of individuals who play the following *stage game* repeatedly: each can *cooperate* (that is, help the other) at a cost $c > 0$ to himself, providing a benefit to the other of $b > c$. Alternatively, each player can *defect*, incurring no cost and providing no benefit. Clearly, both would gain by cooperating in the stage game, each receiving a net gain of $b - c > 0$. However, the structure of the game is that of a *Prisoner's Dilemma*, in which a self-regarding player earns higher payoff by defecting, no matter what his partner does.

The behaviour whereby each individual provides aid as long as this aid has been reciprocated by the other in the previous encounter, is called *tit for tat*. Although termed 'reciprocal altruism' by biologists, this behaviour is self-regarding, because each individual's decisions depend only on the expected net benefit the individual enjoys from the long-term relationship.

On the assumption that after each round of play the interaction will be continued with probability δ, and that players have discount factor d (so $d = 1/(1 + r)$, where r is the rate of time preference), then provided

$$\delta d b > c \tag{1}$$

each individual paired with a tit-for-tat player does better by cooperating (that is, playing tit for tat) rather than by defecting. Thus tit for tat is a best response to itself. To see this, let $\mathbf{v}$ be the present value of cooperating when paired with a tit-for-tat player. The

$$\mathbf{v} = b - c + \delta d \mathbf{v} \tag{2}$$

which gives

$$\mathbf{v} = \frac{b - c}{1 - \delta d} \tag{3}$$

The present value of defecting for ever on a tit-for-tat playing partner is b (the single period gain of b being followed by zero gains in every subsequent period as a result of the tit-for-tat player's defection), so playing tit-for-tat is a best response to itself if and only if $(b - c)(1 - \delta d) > b$, which reduces to (1). Under these conditions unconditional defect is also a best response to itself, so either cooperation or defection can be sustained.

But suppose that, instead of defection for ever, the alternative to tit for tat is for a player to defect for a certain number of rounds, before returning to cooperation on round $k > 0$. The payoff to this strategy against tit for tat is $b - (\delta d)^k c + (\delta d)^{k+1}\mathbf{v}$. This payoff must not be greater than $\mathbf{v}$ if tit for tat is to be a best response to itself. It is

an easy exercise in algebra to show that the inequality

$$\mathbf{v} \geq b - (\delta d)^k c + (\delta d)^{k+1}\mathbf{v}$$

simplifies to (1), no matter what the value of k. A similar argument shows that when (1) holds, defecting for ever (that is, $k = \infty$) does not have a higher payoff than cooperating.

Cooperation Through Reputation Maintenance

Tit for tat takes the form of frequent repetition of the Prisoner's Dilemma stage game inducing a pair of self-regarding individuals to cooperate. In a sizable group, an individual may interact frequently with a large number of partners but infrequently with any single one, say on the average of once every k periods. Players then discount future gains so that a payoff of $\mathbf{v}$ in k periods from now is worth $d^k\mathbf{v}$ now. Then, an argument parallel to that of the previous section shows that cooperating is a best response if and only if

$$\frac{b-c}{1-\delta d^k} > b$$

which reduces to

$$\delta d^k b > c \tag{4}$$

Note that this is the same equation as (1) except that the effective discount factor falls from d to d^k. For sufficiently large k, it will not pay to cooperate. Therefore, the conditions for tit-for-tat reciprocity will not obtain.

But cooperation may be sustained in this situation if each individual keeps a mental model of exactly which group members cooperated in the previous period and which did not. In this case, players may cooperate in order to cultivate a *reputation for cooperation*. When individuals tend to cooperate with others who have a reputation for cooperation, a process called *indirect reciprocity* can sustain cooperation. Let us say that an individual who cooperated in the previous period *in good standing*, and specify that the only way an individual can fall into *bad standing* is by defecting on a partner who is in good standing. Note that an individual can always defect when his partner is in bad standing without losing his good standing status. In this more general setting the tit-for-tat strategy is replaced by the following *standing strategy*: cooperate if and only if your current partner is in good standing, except that, if you accidentally defected the previous period, cooperate this period unconditionally, thereby restoring your status as a member in good standing. This *standing model* is due to Sugden (1986).

Panchanathan and Boyd (2004) have proposed an ingenious deployment of indirect reciprocity, assuming that there is an ongoing dyadic helping game in society based on the indirect reciprocity information and incentive structure, and there is also an n-player public goods game, played relatively infrequently by the same individuals. In the dyadic helping game, two individuals are paired and each member of the pair may confer a benefit b upon his partner at a cost c to himself, an individual remaining in good standing so long as he does not defect on a partner who is in good standing. This random pairing is repeated with probability δ and with discount factor d. In the public goods game, an individual produces a benefit b_g that is shared equally by all the other members, at a cost c_g to himself. The two games are linked by defectors in the public goods game being considered in bad standing at the start of the helping game that directly follows. Then, cooperation can be sustained in both the public goods game and in the dyadic helping game so long as

$$c_g \leq \frac{b(1-\mathcal{E}) - c}{1-\delta d} \tag{5}$$

where ε is the rate at which cooperators unintentionally fail to produce the benefit. Parameters favouring this solution are that the cost c_g of cooperating in the public goods game be low, the factor δd is close to unity, and the net benefit $b(1-\varepsilon) - c$ of cooperating in the reputation-building reciprocity game be large.

The major weakness of the standing model is its demanding informational requirements. Each

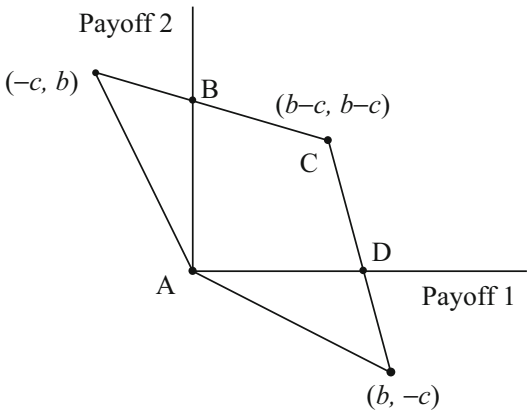
individual must know the current standing of each member of the group, the identity of each member's current partner, and whether each individual cooperated or defected against his current partner. Since dyadic interactions are generally private, and hence are unlikely to be observed by more than a small number of others, errors in determining the standing of individuals may be frequent. This contrasts sharply with the repeated game models of the previous section, which require only that an individual know how many of his current partners defected in the previous period. Especially serious is that warranted noncooperation (because in one's own mental accounting one's partner is in bad standing) may be perceived to be unwarranted defection by some third parties but not by others. This will occur with high frequency if information partially private rather than public (not everyone has the same information). It has been proposed that gossip and other forms of communication can transform private into public information, but how this might occur among self-regarding individuals has not been (and probably cannot be) shown, because in any practical setting individuals may benefit by reporting dishonestly on what they have observed, and self-regarding individuals do not care about the harm to others induced by false information. Under such conditions, disagreements among individuals about who ought to be punished can reach extremely high levels, with the unravelling of cooperation as a result.

In response to this weakness of the standing model, Nowak and Sigmund (1998) developed an indirect reciprocity model which they term *image scoring*. Players in the image scoring need not know the standing of recipients of aid, so the informational requirements of indirect reciprocity are considerably reduced. Nowak and Sigmund show that the strategy of cooperating with others who have cooperated in the past, *independent of the reputation of the cooperator's partner*, is stable against invasion by defectors, and weakly stable against invasion by unconditional cooperators once defectors are eliminated from the population. Leimar and Hammerstein (2001), Panchanathan and Boyd (2003), and Brandt and Sigmund (2004, 2005), explore the applicability of image scoring.

Cooperation in Large Groups of Self-Regarding Individuals

Repeated game theory has extended the above two-player results to a general *n*player stage game, the so-called *public goods game*. In this game each player cooperates at cost $c > 0$, contributing an amount $b > c$ that is shared equally among the other $n - 1$ players. We define the *feasible payoff set* as the set of possible payoffs to the various players, assuming each cooperates with a certain probability, and each player does at least as well as the payoffs obtaining under mutual defection. The set of feasible payoffs for a two-player public goods game is given in Fig. 1 by the four-sided figure ABCD. For the *n*-player game, the figure ABCD is replaced by a similar *n*-dimensional polytope.

Repeated game models have demonstrated the so-called folk theorem, which asserts that any distribution of payoffs to the n players that lies in the feasible payoff set can be supported by an equilibrium in the repeated public goods game, provided the discount factor times the probability of continuation, δd, is sufficiently close to unity. The equilibrium concept employed is a refinement of subgame perfect equilibrium. Significant contributions to this literature include Fudenberg and



Cooperation, Fig. 1 Two-player public goods game

Maskin (1986), assuming perfect information, Fudenberg et al. (1994), assuming imperfect information, so that cooperation is sometimes inaccurately reported as defection, and Sekiguchi (1997), Piccione (2002), Ely and Välimäki (2002), Bhaskar and Obara (2002) and Mailath and Morris (2006), who assume that different players receive different, possibly inaccurate, information concerning the behaviour of the other players.

The folk theorem is an *existence theorem* affirming that any outcome that is a Pareto improvement over universal defection may be supported by a Nash equilibrium, including point C (full cooperation) in the figure and outcomes barely superior to A (universal defection). The theorem is silent on which of this vast number of equilibria is more likely to be observed or how they might be attained. When these issues are addressed two problems are immediately apparent: first, equilibria in the public goods game supported in this manner exhibit very little cooperation if large numbers of individuals are involved or errors in execution and perception are large, and second, the equilibria are not robust because they require some mechanism allowing coordination on highly complex strategies. While such a mechanism could be provided by centralized authority, decentralized mechanisms, as we will see, are not sustainable in a plausible dynamic.

The Dynamics of Cooperation

The first difficulty, the inability to support high levels of cooperation in large groups or with significant behavioural or perceptual noise, stems from the fact that the only way players may punish defectors is *to withdraw their own cooperation*. In the twoperson case, defectors are thus targeted for punishment. But for large n, withdrawal of cooperation to punish a single defector punishes all group members equally, most of whom, in the neighbourhood of a cooperative equilibrium, will be cooperators. Moreover, in large groups, the rate at which erroneous signals are propagated will generally increase with group size, and the larger the group, the larger the fraction of time group members will spend punishing (miscreants and fellow cooperators alike). For instance, suppose the rate at which cooperators accidentally fail to produce b, and hence signal defection, is five per cent. Then, in a group of size two, a perceived defection will occur in about ten per cent of all periods, while in a group of size 20, at least one perceived defection will occur in about 64 per cent of all periods.

As a result of these difficulties, the folk theorem assertion that we can approximate the per-period expected payoff as close to the efficient level (point C in Fig. 1) as desired as long as the discount factor δ is sufficiently close to unity is of little practical relevance. The reason is that as $\delta \to 1$, the current payoff approximates zero, and the expected payoff is deferred to future periods at very little cost, since future returns are discounted at a very low rate. Indeed, with the discount factor δ held constant, the efficiency of cooperation in the Fudenberg, Levine and Maskin model declines at an exponential rate with increasing group size (Bowles and Gintis 2007, ch. 13). Moreover, in an agent-based simulation of the public goods with punishment model, on the assumption of a benefit/cost ratio of $b/c = 2$ (that is, contributing to the public good costs half of the benefit conferred on members of the group) and a discount factor times probability of repetition of $d\delta = 0.96$, even for an error rate as low as $\varepsilon = 0.04$, fewer than half of the members contribute to the public good in groups of size $n = 4$, and less that 20 per cent contribute in groups of size $n = 6$ (Bowles and Gintis 2007, ch. 5).

The second limitation of the folk theorem analysis is that it has not been shown (and probably cannot be shown) that the equilibria supporting cooperation are dynamically robust, that is, asymptotically stable with a large basin of attraction in the relevant dynamic. Equilibria for which this is not the case will seldom be observed because they are unlikely to be attained and if attained unlikely to persist for long.

The Nash equilibrium concept applies when each individual expects all others to play their parts in the equilibrium. But, when there are multiple equilibria, as in the case of the folk theorem,

where there are many possible patterns of response to given pattern of defection, each imposing distinct costs and requiring distinct, possibly stochastic, behaviours on the part of players, there is no single set of beliefs and expectations that group members can settle upon to coordinate their actions (Aumann and Brandenburger 1995).

While game theory does not provide an analysis of how beliefs and expectations are aligned in a manner allowing cooperation to occur, sociologists (Durkheim 1902; Parsons and Shils 1951) and anthropologists (Benedict 1934; Boyd and Richerson 1985; Brown 1991) have found that virtually every society has such processes, and that they are key to understanding strategic interaction. Borrowing a page from sociological theory, we posit that groups may have *focal rules* that are common knowledge among group members. Focal rules could suggest which of a countless number of strategies that could constitute a Nash equilibrium should all individuals adopt them, thereby providing the coordination necessary to support cooperation. These focal rules do not ensure equilibrium, because error, mutation, migration, and other dynamical forces ensure that on average not all individuals conform to the focal rules of the groups to which they belong. Moreover, a group's focal rules are themselves subject to dynamical forces, those producing better outcomes for their members displacing less effective focal rules.

In the case of the repeated public goods game, which is the appropriate model for many forms of large-scale cooperation, Gintis (2007) shows that focal rules capable of supporting the kinds of cooperative equilibria identified by the folk theorem are not evolutionarily stable, meaning that groups whose focal rules support highly cooperative equilibria do worse than groups with less stringent focal rules, and as a result the focal rules necessary for cooperation are eventually eliminated.

The mechanism behind this result can be easily explained. Suppose a large population consists of many smaller groups playing n-person public goods games, with considerable migration across groups, and with the focal rules of successful groups being copied by less successful groups. To maintain a high level of cooperation in a group, focal rules should foster punishing defectors by withdrawing cooperation. However, such punishment is both costly and provides an external benefit to other groups by reducing the frequency of defection-prone individuals who might migrate elsewhere. Hence, groups that 'free ride' by not punishing defectors harshly will support higher payoffs for its members than groups that punish assiduously. Such groups will then be copied by other groups, leading to a secular decline in the frequency of punishment suggested by focal rules in all groups. Thus, suppose that the groups in question were competitive firms whose profits depend on the degree of cooperation among firm members. If all adopted a zero-tolerance rule (all would defect if even a single defection was perceived), then a firm adopting a rule that tolerated a single defection would sustain higher profits and replace the zero-tolerance firms. But this firm would in turn be replaced by a firm adopting a rule that tolerates two defections.

These two problems – the inability to support efficient levels of cooperation in large groups with noisy information, and dynamic instability – have been shown for the case where information is public. Private information, in general the more relevant case, considerably exacerbates these problems.

Cooperation with Other-Regarding Individuals

The models reviewed thus far have assumed that individuals are entirely self-regarding. But cooperation in sizable groups is possible if there exist other-regarding individuals in the form of *strong reciprocators*, who cooperate with one another and punish defectors, even if they sustain net costs. Strong reciprocators are altruistic in the standard sense that they confer benefits on other members of their group (in this case, because their altruistic punishment of defectors sustains cooperation) but would increase their own payoffs by adopting self-regarding behaviours. A model with *social preferences* of this type can explain large-scale decentralized cooperation with noisy

information as long as the information structure is such that defectors expect a level of punishment greater than costs of cooperating.

Cooperation is not a puzzle if a sufficient number of individuals with social preferences are involved. The puzzle that arises is how such altruistic behaviour could have become common, given that bearing costs to support the benefits of others reduces payoffs, and both cultural and genetic updating of behaviours is likely to favour traits with higher payoffs. This evolutionary puzzle applies to strong reciprocity. Since punishment is costly to the individual, and an individual could escape punishment by cooperating, while avoiding the costs of punishment by not punishing, we are obliged to exhibit a mechanism whereby strong reciprocators could proliferate when rare and be sustained in equilibrium, despite their altruistic behaviour.

This is carried out in Sethi and Somanathan (2001), Gintis (2000), Boyd et al. (2003), Gintis (2003) and Bowles and Gintis (2004). The evolutionary viability of other types of altruistic cooperation is demonstrated in Bowles et al. (2003), Boyd et al. (2003), Bergstrom (1995) and Salomonsson and Weibull (2006). The critical condition allowing the evolution of strong reciprocity and other forms of altruistic social preferences is that individuals with social preferences are more likely than random to interact with others with social preferences. Positive assortment arises in these models due to deliberate exclusion of those who have defected in the past (by ostracism, for example), random differences in the composition of groups (due to small group size and limited between-group mobility), limited dispersion of close kin who share common genetic and cultural inheritance, and processes of social learning such as conformism or group level socialization contributing to homogeneity within groups. As in the repeated game models, smaller groups favour cooperation, but in this case for a different reason: positive assortment tends to decline with group size. But the group sizes that sustain the altruistic preferences that support cooperative outcomes in these models are at least an order of magnitude larger than those indicated for the repeated game models studied above.

In sum, we think that other-regarding preferences provide a compelling account of many forms of human cooperation that are not well explained by repeated game models with self-regarding preferences. Moreover, a number of studies have shown that strong reciprocity and other social preferences are a common human behaviour (Fehr and Gächter 2000; Henrich et al. 2005) and could have emerged and been sustained in a gene-culture co-evolutionary dynamic under conditions experienced by ancestral humans (Bowles 2006). The above models also show that strong reciprocity and other social preferences that support cooperation can evolve and persist even when there are many self-regarding players, where group sizes are substantial, and when behavioural or perception errors are significant.

Conclusion: Economics and the Missing Choreographer

The shortcomings of the economic theory of cooperation based on repeated games strikingly replicate those of economists' other main contribution to the study of decentralized cooperation, namely, general equilibrium theory. Both prove the existence of equilibria with socially desirable properties, while leaving the question of how such equilibria are achieved as an afterthought, thereby exhibiting a curious lack of attention to dynamics and out-of equilibrium behaviour. Both purport to model decentralized interactions but on close inspection require a level of coordination that is not explained, but rather posited as a *deus ex machina*. To ensure that only equilibrium trades are executed, general equilibrium theory resorts to a fictive 'auctioneer'. No counterpart to the auctioneer has been made explicit in the repeated-game approach to cooperation. Highly choreographed coordination on complex strategies capable of deterring defection are supposed to materialize quite without the need for a choreographer.

Humans are unique among living organisms in the degree and range of cooperation among large numbers of substantially unrelated individuals.

The global division of labour and exchange, the modern democratic welfare state, and contemporary warfare alike evidence our distinctiveness. These forms of cooperation emerged historically and are today sustained as a result of the interplay of self-regarding and social preferences operating under the influence of group-level institutions of governance and socialization that favour cooperators, in part by protecting them from exploitation by defectors.

The norms and institutions that have accomplished this evolved over millennia through trial and error. Consider how real-world institutions addressed two of the shoals on which the economic models foundered. First, the private nature of information, as we have seen, makes it virtually impossible to coordinate the targeted punishment of miscreants. Converting private information about transgressions into public information that can provide the basis of punishment often involves civil or criminal trials, elaborate processes that rely on commonly agreed upon rules of evidence and ethical norms of appropriate behaviour. Even these complex institutions frequently fail to transform the private protestations of innocence and guilt into common knowledge. Second, as in the standing models with private information, cooperation often unravels when the withdrawal of cooperation by the civic-minded intending to punish a defector is interpreted by others as a violation of a cooperative norm, inviting further defections. In all successful modern societies, this problem was eventually addressed by the creation of a corps of specialists entrusted with carrying out the more severe of society's punishments, whose uniforms conveyed the civic purpose of the punishments they meted out, and whose professional norms, it was hoped, would ensure that the power to punish was not used for personal gain.

Like court proceedings, this institution works imperfectly. It is hardly surprising then that economists have encountered difficulty in devising simple models of how large numbers of self-regarding individuals might sustain cooperation in a truly decentralized setting.

Modelling this complex process is a major challenge of contemporary science. Economic theory, favouring parsimony over realism, has instead sought to explain cooperation without reference to other-regarding preferences and with minimalist or fictive descriptions of social institutions.

C

See Also

- ▶ Agent-Based Models
- ▶ Behavioural Economics and Game Theory
- ▶ Behavioural Game Theory
- ▶ Evolutionary Economics
- ▶ Game Theory
- ▶ Group Selection
- ▶ Public Goods Experiments
- ▶ Repeated Games
- ▶ Social Preferences

Bibliography

Arrow, K.J., and G. Debreu. 1954. Existence of an equilibrium for a competitive economy. *Econometrica* 22: 265–290.

Arrow, K.J., and F. Hahn. 1971. *General competitive analysis*. San Francisco: Holden-Day.

Aumann, R.J., and A. Brandenburger. 1995. Epistemic conditions for Nash equilibrium. *Econometrica* 65: 1161–1180.

Axelrod, R., and W.D. Hamilton. 1981. The evolution of cooperation. *Science* 211: 1390–1396.

Benedict, R. 1934. *Patterns of culture*. Boston: Houghton Mifflin.

Bergstrom, T.C. 1995. On the evolution of altruistic ethical rules for siblings. *American Economic Review* 85: 58–81.

Bhaskar, V., and I. Obara. 2002. Belief-based equilibria the repeated Prisoner's Dilemma with private monitoring. *Journal of Economic Theory* 102: 40–69.

Binmore, K. 1993. *Game theory and the social contract: Playing fair*. Cambridge, MA: MIT Press.

Binmore, K. 1998. *Game theory and the social contract: Just playing*. Cambridge, MA: MIT Press.

Binmore, K.G. 2005. *Natural justice*. Oxford: Oxford University Press.

Blau, P. 1964. *Exchange and power in social life*. New York: Wiley.

Bowles, S. 2006. Group competition, reproductive leveling, and the evolution of human altruism. *Science* 314: 1669–1672.

Bowles, S., and H. Gintis. 2004. The evolution of strong reciprocity: Cooperation in heterogeneous populations. *Theoretical Population Biology* 65: 17–28.

Bowles, S., and H. Gintis. 2007. *A cooperative species: Human reciprocity and its evolution, in preparation*.
Bowles, S., C. Jung-Kyoo, and A. Hopfensitz. 2003. The co-evolution of individual behaviors and social institutions. *Journal of Theoretical Biology* 223: 135–147.
Boyd, R., H. Gintis, S. Bowles, and P.J. Richerson. 2003. Evolution of altruistic punishment. *Proceedings of the National Academy of Sciences* 100: 3531–3535.
Boyd, R., and P.J. Richerson. 1985. *Culture and the evolutionary process*. Chicago: University of Chicago Press.
Brandt, H., and K. Sigmund. 2004. The logic of reprobation: Assessment and action rules for indirect reciprocation. *Journal of Theoretical Biology* 231: 475–486.
Brandt, H., and K. Sigmund. 2005. Indirect reciprocity, image scoring, and moral hazard. *Proceedings of the National Academy of Sciences* 102: 2666–2670.
Brown, D.E. 1991. *Human Universals*. New York: McGraw-Hill.
Debreu, G. 1959. *Theory of value*. New York: Wiley.
Durkheim, E. 1902. *De La Division du Travail Social*, 1967. Paris: Presses Universitaires de France.
Ely, J.C., and J. Välimäki. 2002. A robust folk theorem for the Prisoner's Dilemma. *Journal of Economic Theory* 102: 84–105.
Fehr, E., and S. Gächter. 2000. Cooperation and punishment. *American Economic Review* 90: 980–994.
Fudenberg, D., and E. Maskin. 1986. The Folk Theorem in repeated games with discounting or with incomplete information. *Econometrica* 54: 533–554.
Fudenberg, D., D.K. Levine, and E. Maskin. 1994. The Folk Theorem with imperfect public information. *Econometrica* 62: 997–1039.
Gintis, H. 1976. The nature of the labor exchange and the theory of capitalist production. *Review of Radical Political Economics* 8(2): 36–54.
Gintis, H. 2000. Strong reciprocity and human sociality. *Journal of Theoretical Biology* 206: 169–179.
Gintis, H. 2003. The hitchhiker's guide to altruism: Genes, culture, and the internalization of norms. *Journal of Theoretical Biology* 220: 407–418.
Gintis, H. 2007. *Modeling cooperation with self-regarding agents*. Santa Fe: Santa Fe Institute.
Henrich, J., R. Boyd, S. Bowles, C. Camerer, et al. 2005. Economic man in crosscultural perspective: Behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences* 28: 795–815.
Laffont, J.J. 2000. *Incentives and political economy*. Oxford: Oxford University Press.
Leimar, O., and P. Hammerstein. 2001. Evolution of cooperation through indirect reciprocity. *Proceedings of the Royal Society of London. Series B* 268: 745–753.
Mailath, G.J., and S. Morris. 2006. Coordination failure in repeated games with almost-public monitoring. *Theoretical Economics* 1: 311–340.
Nowak, M.A., and K. Sigmund. 1998. Evolution of indirect reciprocity by image scoring. *Nature* 393: 573–577.
Panchanathan, K., and R. Boyd. 2003. A tale of two defectors: The importance of standing for evolution of indirect reciprocity. *Journal of Theoretical Biology* 224: 115–126.
Panchanathan, K., and R. Boyd. 2004. Indirect reciprocity can stabilize cooperation without the second-order free rider problem. *Nature* 432: 499–502.
Parsons, T., and E. Shils. 1951. *Toward a general theory of action*. Cambridge, MA: Harvard University Press.
Piccione, M. 2002. The repeated Prisoner's Dilemma with imperfect private monitoring. *Journal of Economic Theory* 102: 70–83.
Salomonsson, M., and J. Weibull. 2006. Natural selection and social preferences. *Journal of Theoretical Biology* 239: 79–92.
Sekiguchi, T. 1997. Efficiency in repeated Prisoner's Dilemma with private monitoring. *Journal of Economic Theory* 76: 345–361.
Sethi, R., and E. Somanathan. 2001. Preference evolution and reciprocity. *Journal of Economic Theory* 97: 273–297.
Shubik, M. 1959. *Strategy and market structure: Competition, oligopoly, and the theory of games*. New York: Wiley.
Stiglitz, J. 1987. The causes and consequences of the dependence of quality on price. *Journal of Economic Literature* 25: 1–48.
Sugden, R. 1986. *The economics of rights, co-operation and welfare*. Oxford: Basil Blackwell.
Taylor, M. 1976. *Anarchy and cooperation*. London: Wiley.
Tirole, J. 1988. *The theory of industrial organization*. Cambridge, MA: MIT Press.

Cooperative Equilibrium

A. Mas-Colell

Introduction

The term 'cooperative equilibria' has been imported into economics from game theory. It refers to the equilibria of economic situations modelled by means of cooperative games and solved by appealing to an appropriate cooperative solution concept. The influence is not entirely one way, however. Many game theoretic notions (e.g. Cournot–Nash equilibrium, the Core) are formalizations of pre-existing ideas in economics.

The distinguishing feature of the cooperative approach in game theory and economics is that it

does not attempt to model how a group of economic agents (say a buyer and a seller) may communicate among themselves. The typical starting point is the hypothesis that, in principle, any subgroup of economic agents (or perhaps some distinguished subgroups) has a clear picture of the possibilities of joint action and that its members can communicate freely before the formal play starts. Obviously, what is left out of cooperative theory is very substantial. The justification, or so one hopes, is that the drastic simplification brings to centre stage the implications of actual or potential coalition formation. In their classic book, von Neumann and Morgenstern (1944) already emphasized that the possibility of strategic coalition formation was the key aspect setting apart two from three or more players' games.

The previous remarks emphasize free preplay communication as the essential distinguishing characteristic of cooperative theory. There is a second feature common to most of the literature but which nonetheless may not be intrinsic to the theory (this the future will determine). We refer to the assumed extensive ability of coalitions' players to commit to a course of action once an agreement has been reached.

The remaining exposition is divided in three sections. Sections "The Dominance Approach" and "The Valuation Approach" discuss the two main approaches to cooperative theory (domination and valuation, respectively). Section "Consistency Qualifications" contains qualifications to the domination approach.

An excellent reference for the topic of this entry is Shubik (1983).

The Dominance Approach

Suppose we have N economic agents. Every agent has a strategy set S_i. Denote $S = S_1 \times \ldots \times S_N$ with generic element $S = (S_1, \ldots, \times S_N)$. Given s and a coalition $C \subset N$, the expression s_C denotes the strategies corresponding to members of C. Letting C' be complement of C, the expression $(s_C, s_{C'}$ defines s in the obvious way. For every i there is a utility function $u_i(s)$. If $u = (u_1, \ldots, u_N)$ is an N-list of utilities, expressions such as u_C or $(u_C, u_{C'})$ have the obvious meaning.

Example 1 (Exchange economies): There are N consumers and l desirable goods. Each consumer has a utility function $u_i(x_i)$ and initial endowments ω_i. A strategy of consumer i is an N non-negative vector $S = (S_{i1}, \ldots, \times S_{iN})$ such that $\sum_{j=1}^{N} s_{ij} \leqslant \omega_i$, i.e. s_i is an allocation of the initial endowments of i among the N consumers. Of course, $u_j(s) = u_j\left(\sum_i s_{ij}\right)$.

Example 2 (Public goods): Suppose that to the model of Example 1 we add a public good y with production function $y = F(v)$. Utility functions have the form $u_j(x_j,y)$. A strategy for i is now an $(N + 1)l$ vector $S_i = (S_{i1}, \ldots, S_{i,N+1})$ where $S_{i,N+1}$ is allocated as input to production. We have $u_j(s) = u_j\left[\sum_i s_{ij}\left(\sum_i s_{i,N+1}\right)\right]$.

Example 3 (Exchange with private bads): This is as the first example, except that there is no free disposal, i.e. $\sum_{j=1}^{N} s_{ij} = \omega_i$ for every i. Some of the goods may actually be bads. To be concrete, suppose that $l = 2$, one of the goods is a desirable numéraire and the other is garbage. All consumers are identical and each owns one unit of numéraire and one of garbage (see Shapley and Shubik 1969).

For a strategy profile s to be called a *cooperative equilibrium* we require that there is no coalition C that *dominates* the utility vector $u(S) = (u_1(s),\ldots,u_N(s))$ i.e. that can 'make effective' for its members utility levels $u_{i,\, i \in C}$, such that $u_i > u_i(s)$ for all $i \in C$ Denote by $V(C)$ the utility levels that C can 'make effective' for its members. The precise content of the equilibrium concept depends, of course, on the definition of $V(C)$. I proceed to discuss several possibilities (Aumann 1959, is a key reference for all this).

(A) In line with the idea of Cournot–Nash equilibrium, we could define $V_s(C) = \{u_C : u_C \leqslant u_C(s'_C, s_C)$ for some $S'_C \in S_C\}$, that is, the agents in C take the strategies of C' as fixed. They do not anticipate, so to speak, any retaliatory move. The cooperative solution concept that uses $V_s(C)$ is called *strong Cournot–Nash equilibrium*. It

is very strong indeed. So strong, that it rarely exists. Obviously, this limits the usefulness of the concept. It is immediately obvious that it does not exist for any of the three examples above.

Note that $V_s(C)$ depends on the reference point s. We now go to the other extreme and consider definitions where when a coalition contemplates deviating, it readies itself for a retaliatory behaviour on the part of the complementary coalition; that is, the deviation erases the initial position and is carried out if and only if better levels of utility can be reached, no matter what the agents outside the coalition do. On defining $V(C)$, however, there is an important subtlety. The set $V(C)$ can be defined as either what the members of C cannot be prevented from getting (i.e. the members of C move second) or, more strictly, as what the members of C can guarantee themselves (i.e. they move first). More precisely:

(B) For every C, define:

$$V_\beta(C) = [u_C\text{: for any } s_C \text{ there is an } s_C \text{ such that } u_C \leq u(s_C, s_{C'})].$$

This is what C cannot be prevented from getting. The set of corresponding cooperative equilibria is called the β-core of the game or economy. For any s we have $V_\beta(C) \subset V_s(C)$, and so there is more of a chance for a β-core equilibrium to exist than for a strong Cournot–Nash equilibrium. But there is no general existence theorem. As we shall see, the β-core is non-empty in examples 1 and 2. It is instructive to verify that it is empty in example 3. By symmetry, it is enough to check that the strategies where each agent consumes its own endowment is not an equilibrium. Take the coalition formed by two of the three (identical) agents. As a retaliatory move, the third agent would, at worst, be dumping its unit of garbage on one of the members of the coalition (or perhaps splitting it among them), but the coalition can still be better off than at the initial endowment point by dumping its two units on the third member *and* transferring some money from the nonreceptor to the receptor of outside garbage.

(C) For every C define:

$$V_\alpha(C) = [u_C : \text{ there is } s_C \text{ such that } u_C \leq u_C(s_C, s_{C'}) \text{ for any } s_{C'}].$$

This is what C can guarantee itself of getting. It represents the most pessimistic appraisal of the possibilities of C. The set of corresponding equilibria is called the α-core of the game or economy. For any s we have $V_\alpha(C) \subset V_\beta(C)$ and so there is more of a chance for an α-core equilibrium to exist than for a β-core equilibrium. For the α-core there is a general existence theorem:

Theorem (Scarf 1971): If S is convex, compact and every $u_i(s)$ is continuous and quasiconcave, then the α-core is non-empty.

The conditions of the above theorem are restrictive. Note that the quasiconcavity of u_i is required for the entire s and not only (as for Cournot–Nash equilibrium) for the vector s_i of own strategies. Nonetheless, it is a useful result. It tells us, for instance, that under the standard quasiconcavity hypothesis on utility functions, the α-core is non-empty in each of the three examples above. It will be instructive to verify why the initial endowment allocation is an equilibrium in example 3. In contrast to the β-core situation, a coalition of two members cannot now improve over the initial endowments because they have to move first and therefore cannot know who of the two will receive the outside member's garbage and will need, as a consequence, some extra amount of money.

If, as in examples 1 and 2, there are no bads, the distinction between V_α and V_β disappears. There is a unique way for the members of C' to hurt C, namely withholding its own resources. So in both the α and β senses the set $V(C)$ represents the utility combinations that can be attained by the members of C using only its own resources. This, incidentally, shows that the β-core is non-empty in examples 1 and 2 (since it is equal to the α-core!). There is another approach to existence in the no-bads case. Indeed, a Walrasian equilibrium (in the case of example 2 this takes the guise of a Lindahl equilibrium) is always in this core with no

need of α or β qualification. In the context of example 1, the Core was first defined and exploited by Edgeworth (1881) (*see* "▶ Cores").

Underlying both the α- and the β-core there is a quite pessimistic appraisal on what C' may do if C deviates. The next two remarks discuss, very informally, other, less extreme, possibilities.

(D) In the context of exchange economies (such as example 1) it seems sensible to suppose that a coalition of buyers and sellers in one market may neglect retaliation possibilities in unrelated markets. As it stands in subsections "The Bargaining Set" and "Coalition-Proof Cournot–Nash Equilibrium", it is very difficult for a group of traders to improve, since, so to speak, they have to set up a separate economy covering all markets. See Mas-Colell (1982) for further discussion of this point.

(E) For transferable utility situations (and for purposes more related to the valuation theory to be discussed in section "The Valuation Approach"), Harsanyi (1959), taking inspiration in Nash (1953), proposed that the total utility of the coalition C be defined as $\sum_{i \in C} u_i(\bar{s}_C, \bar{s}_C)$ where $(\bar{s}_C, \bar{s}_{C'})$ are the minimax strategies of the zero sum game between C and C' obtained by letting the payoff of C be $\sum_{i \in C} u_i(s_C, s_{C'}) - \sum_{i \in C'} u_i(s_C, s_{C'})$. Note: if the minimax strategies are not unique, a further qualification is required.

Consistency Qualifications

In this section, several solution concepts are reviewed. Loosely, their common theme is that coalitions look beyond the one-step deviation possibilities.

The Von Neumann–Morgenstern Stable Set Solutions

Suppose that the game is described to us by the sets $V(C)$ that the members of coalitions of C can make effective for themselves. These sets do not depend on any reference combination of strategies. They are constructed from the underlying situation in some of the ways described in section "The Dominance Approach". One says that the N-tuple of utilities $u \in V(N)$ dominates the N-tuple $v \in V(N)$ via coalition C, denoted $u \succ C^{\upsilon}$ if $u_c \in V$. We write $u \succ \upsilon$ if $\bar{u}$ dominates $\bar{v}$ via some coalition. A *core utility computation* is then any maximal element of $\succ$, i.e. any $u \in V(N)$ which is not dominated by any other imputation.

The following paradoxical situation may easily arise. An imputation u is not in the core. Nonetheless, all the members of any coalition that dominates u are treated, at any core imputation, worse than at u (consider for example, the predicament of a Bertrand duopolist at the joint monopoly outcome). If $\succ$ was transitive, then this could not happen, since (continuity complications aside) for any u there would be a core imputation directly dominating u. But $\succ$ is very far from transitive. The approach of von Neumann and Morgenstern consists in focusing on *sets* of imputations K, called *stable sets*, having the properties: (i) if $\upsilon \in K$ then there is no $\upsilon \in K$ that dominates u (internal stability) and (ii) if $u \in K$ then $\upsilon \succ u$ for some $\upsilon \in u$ (external stability). Note that these are the properties that the set of maximal elements of $\succ$ would have if $\succ$ was transitive. The interpretation of K is as a standard of behaviour. If for any reason the imputations of K are regarded as acceptable, then there is an inner consistency to this: drop all the imputations dominated by an acceptable imputation and what you have left is precisely the set of acceptable imputations.

Important as the von Neumann–Morgenstern solution is, its impact in economics has been limited. There is an existence problem, but the main difficulty is that the sets are very hard to analyse.

The Bargaining Set

This solution was proposed by Aumann and Maschler (1964) and is available in several versions. Describing one of them will give the flavour of what is involved. For an imputation u to be disqualified, it will be necessary, but not sufficient, that it be dominated (in the terminology of bargaining set theory: objected to) via some coalition C*. The objection will not 'stick', i.e. throw u out of the negotiation table as a tentative equilibrium, unless it is found justified. The

justifiability criterion is the following: there is no other coalition C^* having a $v_c^* \in K(C^*)$ with the property that $v_i \geqslant u_i$ for every i and which gives to every common member of C and C^* at least as much as they get at the objection. In other words, an objection can be countered if one of the members left out of the objecting coalition can protect themselves in a credible manner (credible in the sense that they can give to any member of C they need, as much as C gives them).

The bargaining set contains the core and, while it is conceptually quite different from a von Neumann–Morgenstern stable set solution, it still does avoid the most myopic features of the core. It is also much easier to analyse than the stable sets, although it is by no means a straightforward tool. But, again, its impact in economics has so far been limited.

A common aspect of stable set and bargaining set theory is that, implicitly or explicitly, a deviating coalition takes into consideration a subsequent, induced move by other coalitions. This is still true for the next two concepts, with one crucial qualification: a deviating coalition only takes into account subsequent moves of its own subcoalitions.

Coalition-Proof Cournot–Nash Equilibrium

This solution concept has been proposed recently by Bernheim et al. (1987). It can be viewed as a self-consistent enlargement of the set of strong Cournot–Nash equilibria. Consider the simplest case, a three-player game. Given a strategy profile $\bar{s}$, which deviations are possible for two players coalitions? If anything, then we are led to strong Cournot–Nash equilibria. But, there is something inconsistent about this. If the strategy profile $\bar{s}$ is not immune to deviations (i.e. there is no commitment at $\bar{s}$), why should the deviation be so? That is, why should it be possible to commit to a deviation? This suggests that the deviation should be required to be immune to further deviations, that is, they should be Cournot–Nash equilibria of the induced two person game (the third player remains put at $\bar{s}$). Obviously, deviating becomes more difficult and the equilibrium set has more of a chance of being non-empty. Unfortunately, there is no general existence theorem. For three-person games, this is precisely the Coalition-Proof Cournot–Nash equilibrium. By recursion, one obtains a definition for any number of players.

The Core

It may be surprising to list the core in a section on concepts that attempts to be less myopic than the core. But, in fact, the core as a set can be made consistent against further deviations by *subcoalitions* of the deviating coalition. Simply make sure always to deviate via coalitions of smallest possible cardinality.

The Valuation Approach

The aim of the valuation approach to games and conflict situations (of which the Shapley value is the central concept) is to associate to every game a reasonable outcome taking into account and compromising among all the conflicting claims. In games, those are expressed by sets $V(C)$ of utility vectors for which C is effective. The criteria of reasonableness are expressed axiomatically. Thus the valuation approach has to be thought of more as input for an arbitrator than as a descriptive theory of equilibrium. Except perhaps for the bargaining set, this point of view is strikingly different to anything discussed so far.

Sometimes the term ‘fair’ is used in connection with the valuation approach. There are at least two reasons to avoid this usage. The first is that the initial position [embodied in the sets $V(C)$] is taken as given. The second is that the fairness of a solution to a game can hardly be judged in isolation, i.e. independently of the position of the players in the overall socioeconomic game.

The valuation of a game will depend on the claims, i.e. on how the sets $V(C)$ are constructed. We saw in section “The Dominance Approach” that there was nothing straightforward about this. We will not repeat it here. It may be worthwhile to observe informally, however, that the valuation approach is altogether less strategic than the dominance one and that a useful way to think of $V(C)$ is as the utility levels the members of C could get if the members of C' did not exist, rather than as what the members of C could get if they go it

alone [in defining $V(C)$ this point of view can make a difference].

Consider first games with transferable utilities (N,υ) where N is a set of players and $\upsilon:2^N \to R$ is a real valued function satisfying $\upsilon(\varphi) = 0$. The restriction of υ to a $C \in N$ is denoted (C, υ). The *Shapley value* is a certain rule that associates to every game (N,v) an imputation $Sh(N,v)$, i.e. $\sum_{i \in N} Sh^i(N,\upsilon) = \upsilon(N)$.

The *Shapley value* was characterized by Shapley (1953) by four axioms that can be informally described as: (i) efficiency, i.e. $Sh(N,v)$ is an imputation, (ii) symmetry, i.e. the particular names of the players do not matter, (iii) linearity over games and (iv) dummy, i.e. a player that contributes nothing to any coalition receives nothing.

There is a simply way to compute the Shapley value. Put $P(\varphi,v) = 0$ and, recursively, associate to every game (N,v) a number $P(N,v)$ such that

$$\sum_{i \in N} [P(N,\upsilon) - P(N/(i),\upsilon)] = \upsilon(N) \qquad (1)$$

That is, the sum of marginal increments of P equals $\upsilon(N)$. This function is called the *potential* and it turns out that the marginal increments of P constitute precisely the Shapley valuations, i.e. $Sh^i(N, \upsilon) - P[N/(i), \upsilon]$ for all (N, υ) and $i \in N$. This is discussed in Hart and Mas-Colell (1985).

The Shapley value for transferable utility games admits several generalizations to the non-transferable utility case [with convex sets $V(C)$]. See Harsanyi (1959), Shapley (1969), and Aumann (1985). Perhaps the most natural, although not necessarily the simpler to work with, was proposed by Harsanyi (1959) and has recently been axiomatized by Hart (1985). For a given game, an Harsanyi value imputation is obtained by rescaling individual utilities so as to guarantee the existence of an N-tuple $u \in V(N)$ satisfying, simultaneously, (i) the convex set $V(N)$ is supported at u by a hyperplane with normal $q = (1,\ldots,1)$, (ii) if a potential P on the set of all games is defined by formula (1) (but replacing '$=v(N)$' by '$\in$ *Bdry.* $V(N)$') then, as before, $u_i = P(N/(i), V)$ for all $i \in N$.

One of the most striking features of the applications of Shapley value theory to economics is that, in economies with many traders, it has turned out to be intimately related to the notion of Walrasian equilibrium. Interestingly, this is in common with the dominance approach. Aumann (1975) is a representative paper of the very extensive literature on the topic.

See Also

- Collusion
- Cores

References

Aumann, R. 1959. Acceptable points in general cooperative n-person games. *Annals of Mathematics Studies Series* 40: 287–324.

Aumann, R. 1975. Values of markets with a continuum of traders. *Econometrica* 43: 611–646.

Aumann, R. 1985. An axiomatization of the non-transferable utility value. *Econometrica* 53: 599–612.

Aumann, R., and M. Maschler. 1964. The bargaining set for cooperative games. In *Advances in game theory*, ed. M. Dresher, L. Shapley, and A.W. Tucker, 443–447. Princeton: Princeton University Press.

Bernheim, B.D., B. Peleg, and M. Whinston. 1987. Coalition-proof Nash equilibria I. Concepts. *Journal of Economic Theory* 42: 1.

Edgeworth, F. 1881. *Mathematical psychics*. London: Kegan Paul.

Harsanyi, J. 1959. Contributions to the theory of games. In *Contributions to the theory of games*, vol. 4, ed. A.W. Tucker and R.D. Luce, 324–356. Princeton: Princeton University Press.

Hart, S. 1985. An axiomatization of Harsanyi non-transferable utility solution. *Econometrica* 53: 1295–1314.

Hart, S. and Mas-Colell, A. 1985. *The potential: a new approach to the value in multiperson allocation problems*. Harvard Discussion Paper 1157.

Mas-Colell, A. 1982. Perfect competition and the core. *Review of Economic Studies* 49: 15–30.

Nash, J. 1953. Two-person cooperative games. *Econometrica* 21: 128–140.

Scarf, H. 1971. On the existence of a cooperative solution for a general class of n-person games. *Journal of Economic Theory* 3: 169–181.

Shapley, L. 1953. A value for n-person games. In *Contributions to the theory of games*, vol. 2, ed. H. Kuhn and A.W. Tucker, 307–317. Princeton: Princeton University Press.

Shapley, L. 1969. Utility comparison and the theory of games. In *La Décision*. Paris: Editions du CNRS, 251–263.
Shapley, L., and M. Shubik. 1969. On the core of an economic system with externalities. *American Economic Review* 59: 678–684.
Shubik, M. 1983. *Game theory in the social sciences*. Cambridge, MA: MIT Press.
von Neumann, J., and O. Morgenstern. 1944. *Theory of games and economic behavior*. Princeton: Princeton University Press.

Cooperative Games

Martin Shubik

The title 'cooperative games' would be better termed games in coalitional form. The theory of games originally developed different conceptual forms, together with their associated solution concepts, namely, games in extensive form, in strategic form, and in coalitional form (von Neumann and Morgenstern 1944). The game in strategic form is sometimes referred to as the game in normal form, while that in coalitional form is also referred to as the game in characteristic form.

The game in extensive form provides a process account of the detail of individual moves and information structure; the tree structure often employed in its description enables the researcher to keep track of the full history of any play of the game. This is useful for the analysis of reasonably well-structured formal process models where the beginning, end and sequencing of moves is well-defined, but is generally not so useful to describe complex, loosely structured social interaction.

A simple example shows the connections among the three representations of a game.

Consider a game with two players where the rules prescribe that Player A moves first. He must decide between two moves. After he has selected a move, Player B is informed and in turn selects between two moves. After B has selected a move the game ends and depending upon the history of the game each player obtains a payoff. Figure 1a shows this game in extensive form. The vertex labelled 0 indicates the starting point of the game. It is also circled to indicate the information structure. Figure 2a shows a game whose only difference from the game in Fig. 1a is that in the latter Player B when called upon to select a move does not know to which of the choice points in his information set the game has progressed. In the game in Fig. 1a, when Player B makes his choice he knows precisely if Player A has selected move 1 or 2. Each vertex of the game is a choice point except the terminal vertices. Several vertices may be enclosed in the same information set. The player who 'owns' a particular information set is unable to distinguish among the choice points in a set. An arc (or branch of a tree) connecting a choice point with another choice point or a terminal point is a move. The moves emanating from any choice point are indexed so that they can be identified.

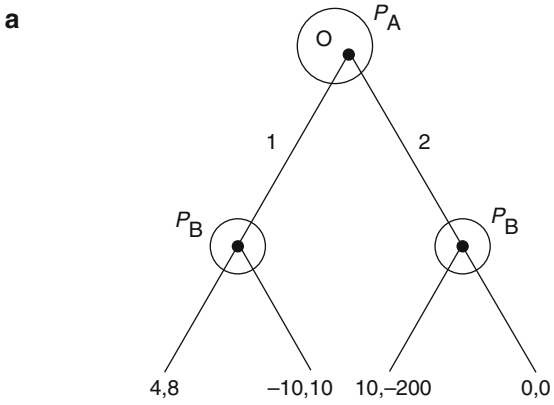
The final nodes at the bottom of the tree are not choice points but points of termination of the game and the numbers displayed indicate the value of the outcome to each player. The first number is the payoff to Player A and the second to Player B.

The extensive form may be reduced to the strategic form by means of strategies. A strategy is a plan covering all contingencies. Figure 1b shows that the moves and strategies for P_A are the same, choose 1 or 2. But P_B has four strategies as he can plan for the contingency that P_B selects 1 or 2. A sample strategy 1, 1; 2, 1 may be read as: 'If P_A selects 1, select 1; if P_A selects 2, select 1'.

The progression from extensive form to strategic form entails loss of fine structure. Details of information are no longer available. There are many extensive forms other than Fig. 1a which are consistent with Fig. 1b.

A further compression of the game representation beyond the strategic form may be called for. At the level of bargaining or diplomacy details of strategy may be of little importance. Instead emphasis is laid upon the value of cooperation. The cooperative or coalitional form represents the game in terms of the jointly optimal outcomes obtainable by every set of players. If payoffs are comparable and side-payments are

Cooperative Games, Fig. 1



b

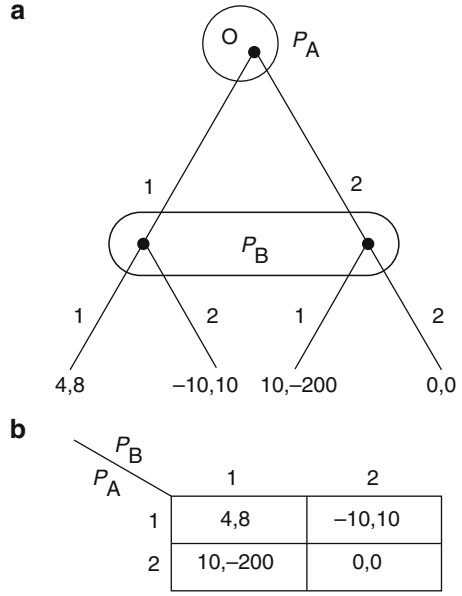
P_A / P_B	1,1;2,1	1,1;2,2	1,2;2,1	1,2;2,2
1	4,8	4,8	–10,10	–10,10
2	10,–200	0,0	10,–200	0,0

possible the gain from cooperation can be represented by a single number. If not then the optimal outcomes attainable by a set S of players will be a Pareto optimal surface in $s = |S|$ dimensions (where $|S|$ is the number of elements in S).

A game in cooperative form with side-payments can be represented by a characteristic function which is a superadditive set function. We use the symbol $\mathbf{\Gamma}(N, \upsilon)$ to stand for a game in coalitional form with a set N of players and a characteristic function v defined on all of the 2^n subsets of N (where $n = |N|$). The condition of superadditivity is a reasonable economic assumption in a transactions cost-free world. $\upsilon(S) + \upsilon(T) \leq \upsilon(S \cup T)$ where $S \cap T = \theta$ states that the amounts obtained by two independent coalitions S and T will be less than or at most equal to the amount that they could obtain by cooperating and acting together.

Returning to Figs. 1b and 2b we can reduce them to coalitional form by specifying how to calculate $\upsilon(\theta), \upsilon(\overline{1}), \upsilon(\overline{2})$ and $\upsilon(\overline{1,2})$. The notation '$\overline{1,2}$' reads as the set consisting of the players whose names are 1 and 2.

Let $\overline{S} = N - S$ be the complement to S. The worst that could happen to S is that $\overline{S}$ acts as a unit



b

P_B / P_A	1	2
1	4,8	–10,10
2	10,–200	0,0

Cooperative Games, Fig. 2

to minimize the joint payoff to S. Applying this highly pessimistic view to the games in Figs. 1b and 2b letting $P_A = 1$ and $P_B = 2$ we obtain the following:

$\upsilon(\theta) = 0$, the coalition of no one obtains nothing, by convention.
$\upsilon(\overline{1}) = 0, \upsilon(\overline{2}) = 0$
$\upsilon(\overline{1,2}) = 12$

Although the extensive and strategic forms of these games differ, they coincide in this coalitional form. More detail has been lost. The coalitional form is symmetric but the underlying games do not appear to be symmetric. The pessimistic way of calculating $\upsilon(S)$ may easily overlook the possibility that it is highly costly for $\overline{S}$ to minimize the payoff to S. Thus it is possible that $\upsilon(S)$ does not reflect the threat structure in the underlying game. Prior to carrying out further game theoretic analysis on a game in characteristic function form the modeller must decide if the characteristic function is an adequate representation of the game. Harsanyi and Selten have suggested a way to evaluate threats (see Shubik 1982).

Applications

Depending upon the application, the extensive, strategic or coalitional forms may be the starting point for analysis. Thus in economic applications involving oligopoly theory one might go from economic data to the strategic form in order to study Cournot-type duopoly. Yet to study the relationship of the Edgeworth contract curve to the price system one can model the coalitional form directly from the economic data without being able even to describe an extensive or strategic form.

In any application, the description of the game in coalitional form is a major step in the specification of the problem. After the coalitional form has been specified a solution is applied to it. There are many solution concepts which have been suggested for games in coalitional form. Among the better known are the core, the value, the nucleolus, the kernel, the bargaining set and the stable set solutions. Only the core and value are noted here (for an exposition of the other see Shubik 1982).

The core of an n-person game in characteristic function form was originally investigated by Gillies and adopted by Shapley as a solution. The value was developed by Shapley (1951) and has been considered in several modifications to account for the presence or absence of threats and sidepayments.

We define $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ where $\alpha_i \geq 0$ for all $i \in N$ and $\Sigma_{i \in N} \alpha_i = \upsilon(N)$ to be an imputation for the game $\Gamma(N, \upsilon)$ It is an individually rational division of the proceeds from total cooperation. The core is the set of imputations such that $\Sigma_{i \in S} \alpha_i \geq \upsilon(S)$ for all $S \subset N$. It is, in some sense, the set of imputations impervious to countervailing power. No subset of players can effectively claim that they could obtain more by acting by themselves. The core may be empty. An exchange economy with the usual Arrow–Debreu assumptions modelled as a game in coalitional form always has a core, and the imputation (or imputations) selected by the competitive equilibria of an exchange economy are always in the core of the associated market game.

The Shapley value is intuitively the average of all marginal contributions that an individual i can make to all coalitions. He developed the explicit formula to calculate the value imputation for any game in coalitional form with sidepayments. It is

$$\phi_i = \sum_{i \in S} \sum_{S \subset N} \frac{(n-s)!(s-l)!}{n!} [\upsilon(S) - \upsilon(S/i)]$$

The term $\upsilon(S) - \upsilon(S/i)$ measures the marginal contribution of i to the coalition S. The remaining terms provide the count of all of the ways the various coalitions involving i can be built up. For exchange economies with many traders a relationship between the competitive equilibria and the value can be established (for further discussion, see Shubik 1984).

Many situations involving voting can be modelled as a game in coalitional form where the characteristic function takes only two values, 0 and 1. Such games are called simple games (Shapley 1962). Shapley and Shubik (1954) suggested the use of the value to provide a power index for committee voting. The basic observation is that the power of a player increases in a nonlinear manner as the number of votes he controls increases. The value applied to a simple game provides an index of this power.

Cooperative games provide a way to carry out an analysis of many problems of interest to the

social sciences without concern for the detail of the structure of process. Von Neumann and Morgenstern aptly noted that the difficulties to be encountered in the development of theories of dynamics in the social sciences were so large that the development of a primarily static theory of games in cooperative form was called for as a first step, bearing in mind that the eventual form of a theory of dynamics might have little resemblance to the statics. Some forty years after their seminal work much still remains to be done in the development of games in coalitional form.

See Also

- ▶ Cooperative Equilibrium
- ▶ Game Theory
- ▶ Nash Equilibrium
- ▶ Non-cooperative Games

Bibliography

Shapley, L.S. 1951. The value of an n-person game. Rand Publication RM-670.
Shapley, L.S. 1962. Simple games: An outline of the descriptive theory. *Behavioral Science* 7: 59–66.
Shapley, L.S., and M. Shubik. 1954. A method for evaluating the distribution of power in a committee system. *The American Political Science Review* 48(3): 787–792.
Shubik, M. 1982. *Game theory in the social sciences*, vol. I. Cambridge, MA: Harvard University Press.
Shubik, M. 1984. *Game theory in the social sciences*, vol. II. Cambridge, MA: MIT Press.
Von Neumann, J., and O. Morgenstern. 1944. *The theory of games and economic behavior*. Princeton: Princeton University Press.

Cooperatives

Ken Coates

It is a very good question to ask why the factory system substituted capitalist for workers' control over the production process. As Andrew Ure remarked,

> To devise and administer a successful code of factory discipline, suited to the necessities of factory diligence, was the Herculean enterprise, the notable achievement of Arkwright . . . it required, in fact, a man of a Napoleon nerve and ambition to subdue the refractory tempers of work-people, accustomed to irregular paroxisms of diligence . . . such was Arkwright (*The Philosophy of Manufactures*, 1835).

Obviously factory discipline was learned reluctantly, and with understandable resentment. Co-operative association did not emerge as a coherent alternative, however, until Robert Owen and his school began to advocate them in the 1820s. Before that date, some friendly societies experimented with cooperative forms of distribution, by bulk purchases of grain and other necessities. Otherwise, those disaffected by the rise of industrial production were more prone, at the beginning, to look to the ownership of land as the basis for communistic experiments. True, there was the pioneering work of the Quaker John Bellers, who proposed a 'college of industry' as early as 1695. But producer co-operatives did not begin to flourish until the 1830s, and even then they had a high failure rate. Industrial disputes, notably the Derby turnouts of 1834, were associated with an insurrectionary idea of co-operation: the Derby workers appealed for help from surrounding towns, not simply to feed those locked out for supporting the Owenite trade union, but also to purchase machinery, so that they could enter into production on their own accord, and begin to construct the co-operative commonwealth.

The defeat of Owenism led to a more gradualist concept of co-operation, although the birth in 1844 of the consumer co-operative movement was far from a devaluation of older communistic ideas. Opening their shop in Toad Lane, the 'Rochdale pioneers' reaffirmed their intentions of raising money in order to embark upon co-operative production. As their initiative spread, it became possible to create a wholesale department servicing several societies, and to open a cornmill and a tobacco factory. Consumer co-operation established the principle of democratic control by its membership, with every member having only one vote, irrespective of the size

of his or her capital investment. Profits were distributed in a dividend on purchases, after collective charges and interest payments had been met.

Because each co-operator votes only once, co-operation establishes a completely different principle of economic administration from that involved in the limited liability company. Retail co-operative societies in Britain frequently grew from the scale at which they could administer one local store to extended chains of shops covering an entire region. Each retail society would be governed by a management committee regularly elected by the membership. Consumer co-operation grew steadily, involving a million people by 1891, five million by 1926, and ten million by 1948. By the 1960s, consumer co-operatives reached a membership of over thirteen million.

With growth, however, there arose problems of involvement. In the original small ventures, all members would be directly involved not only in decision making but in a host of voluntary practical activities. As the movement grew, and recruited professional staff, so its internal democracy became greatly more attenuated. Rates of participation in management meetings declined, reaching very low levels by the middle of the 20th century. A commission under the chairmanship of Hugh Gaitskell and the secretaryship of Anthony Crosland, investigated the decline in co-operative membership involvement. They reported that 'only the few will ever wish to devote their evenings to voluntary public work. In the early days of the co-operative movement, when its total membership was numbered in thousands, almost the entire membership was drawn from among these few . . .. Today, when the movement, has twelve million members, the few have become a small minority . . . the figures of participation have fallen correspondingly.' The assumption of a fixed quota of activists only makes sense, however, on the basis of an assumption of fixed activities in which they might engage. Early co-operation involved a division of tasks among all members, and necessarily engendered high participation ratios. It was professionalization, not simply increase of scale, which changed this situation. As early as 1851, the Rochdale Society had resolved that 'No paid officer be a member of the Board, or a member of the Board a paid servant'. In other words, consumer co-operation in Britain had then established a professional civil service, which was constitutionally excluded from policy control, but increasingly expected to undertake executive management. This is an unreal separation of functions which inevitably eroded the effective powers of individual members. Indeed, when decline set in, apathy, if anything, increased. By 1984 members of co-operatives had declined to 8.5 million, but the 'fixed quota' of activists weighed, if anything, less than before, not more.

Things were quite different in the area of producer co-operation. This grew in labourintensive industries, after the middle of the 19th century, following a successful lobby by Christian Socialists for legislation which could enable them to function. Between 1862 and 1880, 163 producer associations were registered under the Industrial and Provident Societies Acts, of 1852 and subsequently, which had been brought in as a result of careful lobbying by Ludlow, Neale and others. The First International welcomed these new co-operatives, saying of the movement that 'Its great merit is to practically show that . . . the despotic system of the subordination of labour to capital can be superseded by the republican and benificient system of the association of free and equal producers.' A new upsurge in experiments in co-operative workshops followed later, with the labour unrest out of which grew the 'New Unionism' and the 1889 dock strike.

By 1890 Beatrice Webb (Potter) distinguished four classes of producer co-operatives: those modelled on Christian Socialist doctrine, which elected their management committees, and only employed full members; those consisting only of full members who had accepted management by a person (or group) that was (or were) irremovable; those self-governing co-ops which employed outside labour; and those in which outside shareholders supplied most of the capital, but in which workers were encouraged or obliged to take shares, even though they were excluded from the management committee.

While Beatrice Webb documented some exploitative practices in the last three of these

categories, she also pronounced strong judgement on those co-operatives which administered themselves according to strict principles. These, she thought, had a high failure rate, due to their propensity to eat their seed corn and thus fail to make adequate provision for investment. The Webbs' assumption became part of the conventional wisdom about workers' producer co-operatives until the 1970s, when Derek Jones undertook a careful study of the failure rate of producer co-operatives from 1875 onwards and showed that it was not greater than that of small businesses in general. By 1900 there were more than 100 producer co-operatives in Britain, the majority of which had joined forces in a body known as the Co-operative Productive Federation. This organization underwent slow decline, until it listed only 23 societies in the late 1960s. The largest of these, the Leicester Equity Boot and Shoe Manufacturers, employed 1600 people at its peak. Most societies were very much smaller, employing a few dozen.

Producer co-operation underwent a serious revival in the 1970s. Various organizations came into being to argue the case for industrial democracy. A growing discussion on workers' control resulted in the formation of the Institute for Workers' Control (IWC) in 1968, and the extension of debate throughout the organizations of the Labour movement. A specialist body, the Industrial Common Ownership Movement (ICOM) began to organize new co-operatives outside the framework of the Co-operative Productive Federation. This was originally based on initiatives by the Scott Bader Commonwealth, a self-managed chemical company. The industrial policy of the 1974 Labour Government was influenced by the arguments of these new pressure groups, and producer co-operation received major publicity when attempts were made to rescue three failed enterprises from closure by converting them to co-operative management with funding from the Department of Trade and Industry, of which the Secretary of State was then Tony Benn. The 'Benn' co-operatives, as these became known, at Kirkby Manufacturing and Engineering near Liverpool, Triumph–Meriden and the Scottish Daily News, stimulated widespread debate and attracted a number of would-be imitators. A series of 'work ins' and sit-ins, beginning with that at the Upper Clyde shipyards in 1971 had encouraged workers not to accept plant closure when their employers' businesses failed, and some experiments in co-operative production resulted from these struggles. A women's co-operative manufacturing leather goods arose at Fakenham after a factory occupation, and Leadgate Engineering, in the North, followed the same pattern. By 1975, when Imperial typewriters closed their two factories in Hull and Leicester, it had become 'normal' to accept factory occupations as a reflex response to such decisions. The workers in Hull sat in, emblazoning a banner outside the factory, announcing 'We stay in till Benn says when'. But before the typewriter co-operative could be established, Mr Benn was relegated from the Department of Industry to that of Energy, and under his successor no more co-operatives of this kind were to be formed. Soon afterwards, KME and the Scottish Daily News failed for lack of adequate capital. The KME project had been particularly difficult, because it inherited a bizarre product mix, which had itself contributed to the collapse of the original enterprise. The Scottish Daily News came to an end after a series of agreements and disagreements with Robert Maxwell, the publishing entrepreneur, who had been brought in by the workers to assist in the rescue.

But as a result of such colourful events and the persistent lobbying of the industrial democrats, the Labour Party had committed itself to support new co-operatives, and a Co-operative Development agency was established in 1978, with limited funds to promote new organizations. The new agency was not without some opponents in the established co-operative movement, but it was soon to receive a remarkable fillip from the work of Local Government Enterprise Boards, which were established with the onset of slump in the late 1970s in an effort to create employment in local communities. The result was a rapid and spectacular increase in the number of producer organizations. By 1985 there were 750 organizations with an average turnover of £199,000. New co-operatives were being formed all the time, so

that the number of new firms was growing at a rate of 20% per annum.

The recovery of impetus by industrial co-operation already raises important questions for the consumer co-operative movement, and could generate pressures for its reform. It also begins to make possible closer involvement with producer co-operatives in Europe, where they have maintained, in many countries, a consistent strength and vitality. There can be little doubt that, if the upsurge of new co-operative continues, this will bring about-important changes in the political field, as the Labour Movement digests their implications.

References

Bailey, J. 1955. *The British co-operative movement*. London: Hutchinson.

Bradley, K., and A. Gélb. 1983. *Co-operation at work: the Mondragon experience*. London: Heinemann.

Carr-Saunders, A.M., et al. 1938. *Consumers' co-operation in Great Britain*. London: Allen & Unwin.

Coates, K. (ed.). 1976. *The new worker co-operatives*. Nottingham: Spokesman.

Coates, K. 1981. *Work-ins, sit-ins and industrial democracy*. Nottingham: Spokesman.

Cole, G.D.H. 1944. *A century of co-operation*. London: Allen & Unwin.

Derrick, P., and J.F. Phipps. 1969. *Co-ownership, co-operation and control*. London: Longmans.

Eccles, T. 1981. *Under new management*. London: Pan.

Garnett, R.G. 1972. *Co-operation and the Owenite socialist communities in Britain*. Manchester: Manchester University Press.

Greater London Enterprise Board. 1984. *A strategy for co-operation: Worker co-ops in London*. London: GLEB.

ICOM. *The co-operative way: Worker co-ops in France, Spain and Eastern Europe*. London: ICOM Coop Publications.

Jones, D. 1976. British producer co-operatives. In Coates (1976).

Labour Party Finance and Industry Group. 1983. *Towards common ownership*. London: Labour Party.

Marglin, S.A. 1976. What do bosses do? In *The division of labour*, ed. A. Gorz. Brighton: Harvester Press.

Ostergaard, G.N., and A.H. Halsey. 1965. *Power in co-operatives*. Oxford: Blackwell.

Potter, B. 1907. *The co-operative movement in Great Britain*. London: Swan Sonnenschein.

The co-operative directory, (yearly). Manchester: Co-operative Union.

Thomas, H.B., and C. Logan. 1982. *Mondragon: An economic analysis*. London: Allen & Unwin.

Thornley, J. 1981. *Workers' co-operatives: Jobs and dreams*. London: Heinemann.

Ure, A. 1835. *The philosophy of manufactures*. Edinburgh: Charles Knight.

Coordination Problems and Communication

Jack Ochs

Abstract

Coordination problems arise when a game has multiple Nash equilibria and all players have a common interest in avoiding a non-equilibrium state. To achieve an equilibrium state, agents must come to understand one another's intentions. Communication can facilitate this understanding under some, but not all, circumstances. In the absence of communication among agents, coordination may also sometimes be achieved with the aid of extrinsic signals that have come to be associated with the actions of others. In some settings, past actions themselves serve as precedents, without the benefit of any communication.

Keywords

Cheap talk; Communication; Coordination equilibrium; Coordination problems; Extensive form games; Nash equilibrium; Observability; Prisoners' Dilemma; Signalling; Sunspot equilibrium

JEL Classifications

C9

Lewis (1969) defined a *coordination equilibrium* as a Nash equilibrium in which no agent would be better off if any other agent had chosen a different action. When there are multiple coordination equilibria, agents face an obvious coordination problem. The resolution of coordination problems

rests upon individuals coming to understand the intentions of one another. The most explicit way of developing this understanding is for the individuals to communicate with one another. Common knowledge of a language must precede communication. Even with common knowledge of a language, individuals may not be bound to do what they say they will do. In such circumstances, talk is 'cheap'.

When will the receiver, having received a message from a sender, behave differently from how the receiver would have behaved if no message had been sent? According to Farrell and Rabin (1996) *highly credible* messages will not be ignored. A message that signals an intention to take action X is highly credible if it satisfies two conditions: it is (*a*) *self-signalling* and (*b*) *self-committing*. A message that the sender is taking action X is self-signalling if, and only if, it is both true and it is in the sender's interest to have it believed to be true. A message is self-committing if a belief by the receiver that the message is true creates an incentive for the sender to do what the sender said he or she would do. A message that is self-committing, if believed, will lead to an outcome that is a Nash equilibrium. A message can be self-committing without being self-signalling. For example, in the classic game of Chicken, if one player announces that he will be Passive, that message is self-committing since, if it is believed by the receiver then the receiver's best response is to be Aggressive, and the best response of the sender to the receiver's aggression is to be Passive. However, the sender would prefer to have the receiver believe that the sender will play Aggression. So the message, 'I intend to play Passive', is not selfsignalling because it is not in the interest of the sender to have the receiver believe it is true.

A message is *cheap talk* if the sender is not bound to do what the message says. Crawford (1998) provides a survey of a number of cheap talk experiments. In experiments with structured communication, either only one player may send a message (one-sided communication) or more than one player can send a message. When the payoff functions of the players are symmetrical, one-sided communication breaks the symmetry of the game without communication. This is sufficient to allow a very high level of coordination. Indeed, in such games one-sided communication is much more effective in promoting coordination that is simultaneous, two-sided communication. This suggests that, when payoff functions are symmetric but players have different preference orderings over equilibria, as in the Battle of the Sexes, the principal impact of one-sided communication is to create an extensive form game in which the symmetry is broken by designating one player as the first mover. In games with Pareto-ordered equilibria communication is not needed to break symmetry, but may be effective in reducing uncertainty about the intentions or, in Crawford's terms, to give 'reassurance'. Empirically this 'reassurance' appears to be most effective in achieving coordination on the Pareto-dominant equilibrium when communication is two-sided, but even one-sided communication has a positive effect on the likelihood of achieving the Pareto-optimal outcome. Furthermore, this effect has been found to be greater when a message was self-signalling than when such a message was only self-committing.

When there are multiple players each player must be interested in, and possibly condition his actions on, the entire message profile. Therefore, the concepts of selfsignalling and self-committing messages may not have much meaning in this context. Nevertheless, there is some evidence that costless pre-play communication can help groups whose members repeatedly interact to achieve more efficient outcomes than is attainable without such communication (Blume and Ortmann 2007).

A signal that is commonly observed may be used to coordinate actions even if the signal does not emanate from any of the players. Traffic signals play this role. We do as these signals say we should do because we believe that others will also do what the signals say they should do. This belief is reinforced by experience, so doing as the signals suggest has simply become a convention that is adopted by drivers. While this convention is backed by law, there is good reason to believe that it is so ingrained in people's expectations that they would continue to act as the signals

suggest even in the absence of any law. Can signals be effective in coordinating actions when the signals are not sent by any of the players and do not themselves have any payoff consequences? Van Huyck et al. (1992) found that, when a game has multiple coordination equilibria, all of which yield the same payoff, a signal from an outside 'moderator' that specifically says 'play a particular equilibrium' produces a very high degree of coordination on the suggested equilibrium, even though absent any signal there is a high frequency of coordination failure. However, in games where the equilibria are Pareto ordered the introduction of a recommendation to play any equilibrium other than the payoff-dominant equilibrium significantly reduces the degree of coordination that is achieved. The authors also found that when there was an equilibrium that provided equal payoff a recommendation to play an equilibrium with unequal payoffs had little influence on how the game was played. Evidently some features, such as symmetry, may be sufficiently strong focal points that the introduction of extrinsic signals may have little influence. Similarly, some features of a game may make some coordination equilibria, once achieved by repeated interaction, exceedingly difficult to displace through the introduction of communication, even if everyone would gain by moving to another coordination equilibrium (Cooper 2006).

A 'sunspot' is a commonly observable event that may have been correlated in the past with different outcomes. For example, published forecasts may have this property. When agents coordinate their actions on a 'sunspot' the resulting equilibrium is called a 'sunspot equilibrium'. Marimon et al. (1993) devised an experiment to see whether they could generate a sunspot equilibrium where prices fluctuate with an extrinsic signal even though the fundamental parameter values remained fixed. During a 'training interval', the colour of a blinking light on a screen was perfectly correlated with a change in a parameter that induced changes in equilibrium prices. After this 'training period' the parameter value was fixed, but the signal continued to vary according to the same process. Prices continued to be volatile but there was little evidence that the variation in the sunspot variable had any effect on the observed price volatility. Duffy and Fisher (2005), using a quite different design, were able to induce sunspot equilibria under restricted conditions. They found that the semantics of the sunspot variable mattered. There were two fundamental equilibria in their design. One equilibrium had a high price, the other a low price. When the sunspot message was either 'high' or 'low' the outcomes of the actions were sometimes correlated with the message. But when the message was either 'sunshine' or 'rain' this correlation was never observed. Evidently, correlation of expectations with the signal depends upon how confident people are that everyone is interpreting the signal in the same way. They also found that information that is generated by observable actions subsequent to the observation of the signal itself tends to diminish the focal power of the signal.

Sometimes actions might 'speak' louder than words. In a Prisoners' Dilemma game the cooperative outcome is not a Nash equilibrium, but it does Pareto-dominate the Nash equilibrium. Since non-cooperation is a dominant strategy a message that one intends to play 'Cooperate' is neither self-committing nor self-signalling. Nevertheless, Duffy and Feltovich (2002) found that when this message was sent it tended to be truthful and also tended to induce a cooperative response. Similarly, when their past actions with other players were observable, subjects were more likely to cooperate than if neither communication nor observability was possible. Furthermore, observation increased the frequency of cooperative choices by more than cheap talk. This suggests that observability of past actions may sometimes be more effective than mere words in helping people achieve a good outcome.

See Also

- ▶ Cheap Talk
- ▶ Experimental Economics
- ▶ Game Theory

Bibliography

Blume, A., and A. Ortmann. 2007. The effects of costless pre-play communication: Experimental evidence from games with Pareto-ranked equilibria. *Journal of Economic Theory* 132: 274–290.

Cooper, D. 2006. Are experienced managers experts at overcoming coordination failure? *Advances in Economic Analysis & Policy* 6(2), Article 6.

Crawford, V. 1998. A survey of experiments on communication via cheap talk. *Journal of Economic Theory* 78: 286–298.

Duffy, J., and N. Feltovich. 2002. Do actions speak louder than words? An experimental comparison of observation and cheap talk. *Games and Economic Behavior* 39: 1–27.

Duffy, J., and E. Fisher. 2005. Sunspots in the laboratory. *American Economic Review* 95: 510–529.

Farrell, J., and M. Rabin. 1996. Cheap talk. *Journal of Economic Perspectives* 10(3): 103–118.

Lewis, D. 1969. *Convention: A philosophical study*. Cambridge, MA: Harvard University Press.

Marimon, R., S. Spear, and S. Sunder. 1993. Expectationally driven market volatility: An experimental study. *Journal of Economic Theory* 61: 74–103.

Van Huyck, J., A. Gillette, and R. Battalio. 1992. Credible assignments in coordination games. *Games and Economic Behavior* 4: 606–626.

Copland, Douglas Berry (1894–1971)

M. Harper

Copland was born at St Andrews, New Zealand, in 1894 and died at Kilmore, Australia, in 1971. Australia's most public applied economist from the 1920s to 1960, he pioneered opportunities for professional economists. He was the first occupant of positions such as Professor of Economics at the University of Tasmania (1920–24), Professor of Commerce at the University of Melbourne (1924–44), President of the Economic Society of Australia and New Zealand (1925), chief editor of the *Economic Record* (1924–45), Australian/New Zealand representative for the Social Sciences Division of the Rockefeller Foundation (1925–54), Vice-Chancellor of the Australian National University (1948–53), Principal, Australian Administrative Staff College (1956–60), and Chairman, Committee for the Economic Development of Australia (1960–66).

Copland was particularly interested in monetary and capital flows and their relation to prices, business cycles and economic development. Stressing Australia's world position as a small, dependent, primary-producing country, his policy advice was often controversial. In 1929 he contributed to the Australian case for tariff protection. During the 1930s depression, he advocated the 'middle way' towards recovery – a policy-mix of deflationary cost, wage and fiscal measures, with reflationary exchange depreciation, expansionary monetary policy and tariff protection. In the 1950s he recommended that Australia avoid restrictions of the sterling area by pursuing a policy of rapid development based on dollar borrowings.

Copland's experiences as economic adviser to governments, Commonwealth Prices Commissioner (1939–45), Australian Minister to China (1946–8) and President of the Economic and Social Council of the United Nations (1955) while High Commissioner to Canada, led to publications on the parameters and mechanisms of economic control, especially within group frameworks – the Australian Commonwealth, the British Commonwealth and international organizations.

Selected Works

1920. *Wheat production in New Zealand: A study in the economics of New Zealand agriculture*. Auckland: Whitcombe & Tombs.

1929. (With J.B. Brigden, E.C. Dyason, L.F. Giblin and C.H. Wickens). *The Australian tariff: An economic enquiry*. Melbourne: Melbourne University Press (Economic Series No. 6).

1930. *Credit and currency control: With special reference to Australia*. Melbourne: Melbourne University Press (Economic Series No. 9).

1934. *Australia in the world crisis 1929–33*. Cambridge: Cambridge University Press.

1945. *The road to high employment: Administrative controls in a free economy*. Sydney: Angus and Robertson.
1953. *Problems of the sterling area with special reference to Australia*. Essays in International Finance No. 17, September 1953. Princeton: Princeton University International Finance Section.
For a bibliography of academic works, see *Economic Record*, March 1960, 173–178.

Copulas

Pravin K. Trivedi

Abstract

Copulas are functional forms that parameterize the joint distribution of random variables based on their stated marginal distributions and a dependence parameter. The approach is based on Sklar's theorem. Copulas provide a general method for modelling dependence between random variables that may exhibit asymmetric dependence, which is often inadequately captured by measures of linear dependence. Copulas are often generated by using mixtures and convex sums. Although a bivariate distribution is the most commonly encountered specification, higher dimensional joint distributions can also be generated.

Keywords

Clayton copula; Copulas; Cumulative distribution functions; GARCH effects; Gaussian copula; Gumbel copula; Marginal distributions; Selection models; Sklar, A.; Sklar's theorem; Tail dependence

JEL Classifications

C1; C51

Sklar introduced copulas in 1959 (Sklar 1973, 1996). Concisely stated, copulas are functions that connect multivariate distributions to their one-dimensional margins. If F is an m-dimensional continuous cumulative distribution function (CDF) with one-dimensional margins $F_1, \ldots, F_m$, then there exists an m-dimensional unique copula C such that $F(x_1, \ldots, x_m) = C(F_1(x_1), \ldots, F_m(x_m))$. In general, marginal distributions alone cannot determine the joint distributions.

Copulas are useful because, first, they represent a method for deriving joint distributions given the fixed marginals, even when marginals belong to different parametric families of distributions; second, in a bivariate context copulas can be used to define nonparametric measures of dependence for pairs of random variables that can capture asymmetric (tail) dependence as well as correlation or linear association.

Copulas and Dependence

We begin with Sklar's theorem. An m-copula is an m-dimensional CDF whose support is contained in $[0,1]^m$ and whose one-dimensional margins are uniform on [0,1]. In other words, an m-copula is an m-dimensional distribution function with all m univariate margins being $U(0,1)$. To see the relationship between distribution functions and copulas, consider a continuous m-variate distribution function $F(y_1, \ldots y_m)$ with univariate marginal distributions $F_1(y_1), \ldots, F_m(y_m)$ and inverse probability transforms (quantile functions) $F_1^{-1}, \ldots, F_m^{-1}$. Then $y_1 = F_1^{-1}(u_1) \sim F_1, \ldots, y_m = F_m^{-1}(u_m) \sim F_m$ where $u_1, \ldots, u_m$ are uniformly distributed variates. Copulas are expressed in terms of marginal CDFs. The transforms of uniform variates are distributed as $F_i (i = 1, \ldots, m)$. Hence

$$F(y_1, \ldots, y_m) = F\left(F_1^{-1}(u_1), \ldots, F_m^{-1}(u_m)\right) = C(u_1, \ldots, u_m), \quad (1)$$

is the unique copula associated with the distribution function. The copula parameterizes a multivariate distribution in terms of its marginals. For an m-variate distribution F, the copula satisfies

$$F(y_1,\ldots,y_m) = C(F_1(y_1),\ldots,F_m(y_m);\theta), \quad (2)$$

where θ is usually a scalar-valued dependence parameter. For many empirical applications, the dependence parameter is the main focus of estimation. Because the marginal distributions may come from different families, copulas are a 'recipe' for generating joint distributions by combining given marginal distributions using a known copula. This construction allows researchers to consider marginal distributions and dependence as two separate, but related, issues.

The functional form of a copula places restrictions on the dependence structure; for example, it may support only positive dependence. Therefore, a pivotal modelling problem is to choose a copula that adequately captures dependence structures of the data without sacrificing attractive properties of the marginals. Copulas are multivariate distribution functions, hence Fre: chet bounds apply. A copula may impose restrictions such that the full coverage between the bounds is not attained.

An important advantage of copulas is that they generate more general measures of dependence than the correlation coefficient. Correlation is a symmetric measure of linear dependence, bounded between + 1 and −1 and invariant with respect to only linear transformations of the variables. By contrast, copulas have an attractive invariance property: the dependence captured by a copula is invariant with respect to increasing and continuous transformations of the marginal distributions. The same copula may be used for, say, the joint distribution of (Y_1, Y_2) as ($ln\ Y_1$,$ln\ Y_2$).

Measures of dependence based on concordance, such as Spearmans's rank correlation (ρ) and Kendall's τ, overcome limitations of the correlation coefficient. In some cases the concordance between extreme (tail) values of random variables is of interest. For example, one may be interested in the probability of the event that stock indexes in two countries exceed (or fall below) given levels. This requires a dependence measure for upper and lower tails of the distribution, rather than a linear correlation measure. Measures of lower and upper tail dependence can be readily derived for a stated copula. The copula dependence parameter θ can be converted to measures of concordance such as Spearman's ρ and Kendall's τ (Nelsen 1999).

C

Examples

Nelsen (1999) and Joe (1997) catalogue many functional forms for copulas. Particularly important is the Archimedean class. Bivariate Archimedean copulas take the general symmetric form

$$C(u_1,u_2;\theta) = \phi^{-1}(\phi(u_1)+\phi(u_2)), \quad (3)$$

where the generator function $\phi(\cdot)$ is a convex decreasing function; for example, $\phi(t) = -\ln(t)$. The dependence parameter θ in imbedded in the functional form of the generator.

The *Clayton copula*, a member of the Archimedean class, takes the form

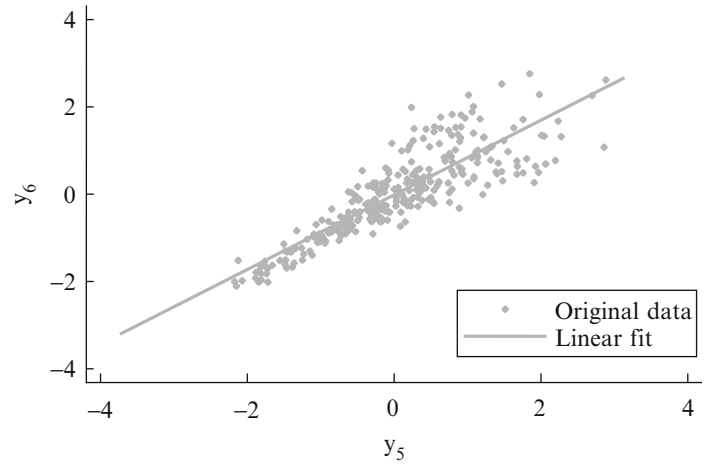
$$C(u_1,u_2;\theta) = \left(u_1^{-\theta}+u_2^{-\theta}-1\right)^{-1/\theta} \quad (4)$$

with the dependence parameter θ restricted on the region $(0,\infty)$. As θ approaches zero, the marginals become independent. The Clayton copula cannot account for negative dependence. It has been used to study correlated risks because it exhibits strong left tail dependence and relatively weak right tail dependence.

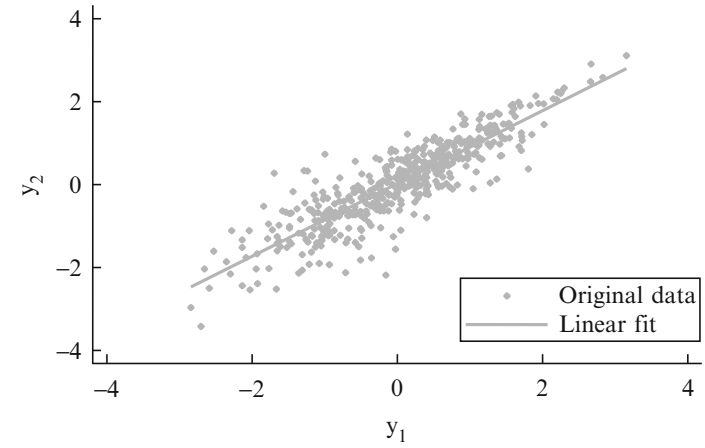
The *Gumbel copula* is another member of the Archimedean class and takes the form

$$C(u_1,u_2;\theta) = \exp\left(-\left(\tilde{u}_1^{\theta}+\tilde{u}_2^{\theta}\right)^{1/\theta}\right) \quad (5)$$

where $\tilde{u}_j = -\log u_j$. The dependence parameter is restricted to the interval $[1,\infty)$. Like the Clayton copula, Gumbel does not allow negative dependence, but in contrast it exhibits strong right tail dependence and relatively weak left tail dependence. If outcomes are strongly correlated at high values but less correlated at low values, then the Gumbel copula is an appropriate choice.



Copulas, Fig. 1 Sample from Clayton copula, theta = 4.67



Copulas, Fig. 2 Sample from Gumbel copula, theta = 3.3

The (non-Archimedean) *Gaussian copula* takes the form

$$C(u_1, u_2; \theta) = \Phi_G\left(\Phi^{-1}(u_1), \Phi^{-1}(u_2); \theta\right), \quad (6)$$

where Φ is the CDF of the standard normal distribution, and $\Phi_G(u_1, u_2)$ is the standard bivariate normal distribution with correlation parameter θ restricted to the interval $(-1, 1)$. This copula allows equal degrees of positive and negative dependence.

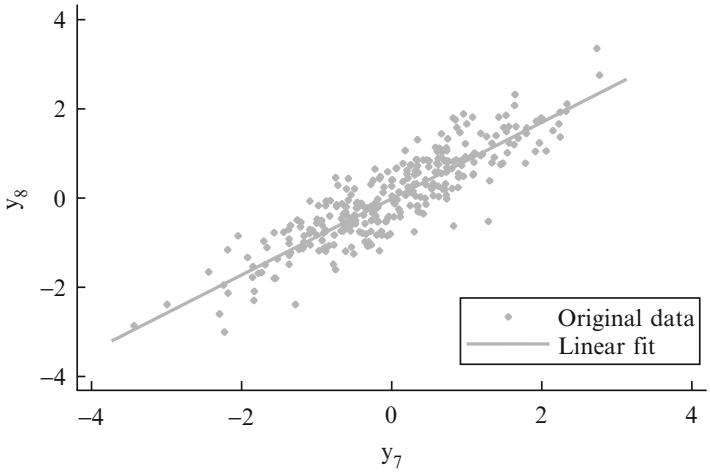
Figures 1, 2, and 3 illustrate lower and upper tail dependence using three samples generated using Monte Carlo draws from the above three copulas. The samples have comparable degrees of linear dependence but different tail dependence properties.

Estimation and Applications

In some applications it would natural to parameterize the marginals in terms of a regression function with covariates z, that is, $u_j = F(y_j|z_j; \beta_j)$, where z_j is a vector of covariates. Then the bivariate copula takes the form $C(y_1, y_2|\, z_1, z_2, \beta_1, \beta_2, \theta) = C(F(y_1|\, z_1, \beta_1), F(y_2|\, z_2, \beta_2), \theta)$. The copula density, defined as

$$\frac{d}{dy_2 dy_1} C(F_1(\cdot), F_2(\cdot); \theta) = C_{12}(F_1(\cdot), F_2(\cdot), f_1(\cdot) f_2(\cdot), \quad (7)$$

$f_j(\cdot) = \partial F_j(\cdot)/\partial y_j$, can be used to build the likelihood, which can be maximized simultaneously



Copulas, Fig. 3 Sample from Gaussian copula, theta = .89

with respect to all unknown parameters. Alternatively, the marginal densities can be estimated first, either parametrically or nonparametrically, and then the likelihood can be maximized with respect to θ only at the second stage.

Multivariate models of survival data pioneered the application of copulas. Econometric applications are more recent, but growing rapidly. There are numerous time series and financial market applications of copulas (Cherubini et al. 2004). Few models in this literature include regressors. Other areas of applications include volatility and exchange rate modelling where GARCH effects and tail dependence are expected (Patton 2006). Selection models provide leading examples of microeconometric applications of copulas (Smith 2003; Zimmer and Trivedi 2006).

See Also

- ▶ Seemingly Unrelated Regressions
- ▶ Simultaneous Equations Models

Bibliography

Cherubini, U., E. Luciano, and W. Vecchiato. 2004. *Copula methods in finance*. New York: John Wiley.

Joe, H. 1997. *Multivariate models and dependence concepts*. London: Chapman and Hall.

Nelsen, R. 1999. *An introduction to copulas*. New York: Springer.

Patton, A. 2006. Estimation of multivariate models for time series of possibly different lengths. *Journal of Applied Econometrics* 21: 147–173.

Sklar, A. 1973. Random variables, joint distributions, and copulas. *Kybernetica* 9: 449–460.

Sklar, A. 1996. Random variables, distribution functions, and copulas – a personal look backward and forward. In *Distributions with fixed marginals and related topics*, ed. L. Ruschendorf, B. Schweizer, and M. Taylor. Hayward: Institute of Mathematic Statistics.

Smith, M. 2003. Modeling selectivity using Archimedean copulas. *Econometrics Journal* 6: 99–123.

Zimmer, D., and P. Trivedi. 2006. Using trivariate copulas to model sample selection and treatment effects: Application to family health care demand. *Journal of Business and Economic Statistics* 24: 63–76.

Core Convergence

Robert M. Anderson

Abstract

The core of an economy is the set of all economic outcomes that cannot be 'blocked' by any group of individuals; it is an institution-free concept. A Walrasian equilibrium is an economic outcome based on the institution of market-clearing via prices: each individual consumes his or her demand, taking prices as given, and the demand for each good equals the

supply of that good. Core convergence asserts that, for sufficiently large economies, every core allocation approximately satisfies the definition of Walrasian equilibrium; it is an important test of the price-taking assumption inherent in the definition of Walrasian equilibrium.

Keywords

Convexity; Cooperative game theory (core); Core convergence; Core; First Welfare Theorem; Edgeworth, F. Y; Second Welfare Theorem; Separating hyperplane th; Shapley–Folkman th; Walrasian equilibrium

JEL Classification

C7; D5

The core of an economy, first defined by Edgeworth (1881), is the set of all economic outcomes such that no group of individuals ('coalition') can make each of its members better off ('improve on' or 'block' the outcome), using only the resources available to the group. (A common mistake is to ask, in reference to a particular core allocation, 'what coalition(s) have formed?' An allocation is in the core precisely when no coalition can improve on it, and a core allocation does not identify an associated coalition or coalitions. It is when an allocation is *not* in the core that one can identify one or more coalitions that are associated with it, because they can improve on it and thus demonstrate that the coalition is not in the core.)

The most important reason for studying the core is the light it sheds on Walrasian equilibrium, introduced by Walras (1874). While the notion of Walrasian equilibrium is based entirely on the institution of trading via prices, and assumes that individuals take prices as given, the definition of the core is completely institution-free; this is one of its major virtues.

The core has both normative and positive significance apart from its relationship to Walrasian equilibrium. Normatively, if one accepts the distribution of the economy's initial resources as equitable, then any allocation outside the core is unfair to at least one coalition. Regardless of whether the distribution of initial resources is equitable, it would be surprising to find the economy settling on an allocation outside the core, since that would indicate there is a coalition which *could* have made each of its members better off, using only its own resources, but for some reason has failed to coalesce and do so; this is the positive significance.

While there has been much work on the cores of production economies, the bulk of the work on the core has been carried out in exchange economies, in which trading and consuming are the only economic activities. In part, this is because there are a number of competing definitions of the core in production economies, based on how the ownership of the production technology is assigned to individuals and groups. For simplicity, we shall focus our attention on exchange economies.

Walrasian equilibrium is an economic equilibrium notion based on market clearing, mediated by prices. Consumers choose the consumption vector which maximizes utility over their budget sets; firms choose production plans which maximize profit. Critically, it is assumed that individuals and firms take prices as given, without taking into account any ability they may have to influence those prices through their actions. A price vector is a Walrasian equilibrium price if the choices made by individuals and firms, taking prices as given, are consistent in the sense that market supply equals market demand. A Walrasian allocation is the vector of individual consumptions and firm productions generated by a Walrasian equilibrium price. A Walrasian equilibrium is a pair consisting of a Walrasian equilibrium price and its associated Walrasian allocation.

An income transfer is a vector which assigns to each agent a real number, and which satisfies budget balance: the sum of the numbers is zero. An allocation is a Walrasian equilibrium with transfers if there is an income transfer and a price vector such that the demand of each agent, given the prices and the budget of the agent, taking into account the agent's endowment of goods and income transfers, just equals the individual's consumption at the allocation.

The First and Second Welfare Theorems are two of the most important results concerning Walrasian

equilibrium. Recall that, in an exchange economy, an allocation is Pareto optimal if there is no reallocation of consumption which makes every agent better off. In other words, the coalition consisting of all agents (coalition of the whole) cannot improve upon the allocation. Thus, it is clear that every core allocation is Pareto optimal.

The First Welfare Theorem asserts that every Walrasian allocation with transfers is Pareto optimal. A slight modification of the proof suffices to show that every Walrasian allocation lies in the core. (Note that it is *not* true that every Walrasian allocation with transfers lies in the core. The income transfers allow us to move consumption among agents. For example, consider the allocation which gives the entire social consumption to a single agent. If we choose a price vector which supports that agent's preference at the social consumption, then there is an income transfer that makes this allocation a Walrasian allocation with transfers. But this allocation will rarely lie in the core, since the coalition consisting of all the other agents will generally be able to improve on it.) This is an important strengthening of the First Welfare Theorem, which has both positive and normative significance. On the positive side, it is a strong stability property of Walrasian equilibrium, since it asserts that no group of individuals would choose to upset the equilibrium by recontracting among themselves, making it more plausible that we will see Walrasian equilibrium arise in real economies. On the normative side, if we accept the distribution of initial endowments as equitable, it tells us that Walrasian allocations are fair to all groups in the economy.

The Second Welfare Theorem asserts that, in an exchange economy with standard assumptions on preferences (convexity is the crucial assumption), every Pareto optimal allocation is a Walrasian equilibrium with transfers. Note that while the definition of Pareto optimality makes no mention of prices, the Second Welfare Theorem asserts that every Pareto optimal allocation is closely associated to a price vector. The price vector appears magically; mathematically, this is a consequence of the separating hyperplane theorem, for which convexity is a critical assumption. As noted above, the most important use of the core is as a test of the price-taking assumption inherent in the definition of Walrasian equilibrium; a number of other tests have been proposed, but *core convergence* is the most commonly used. Core convergence is closely analogous to the conclusion of the Second Welfare Theorem. The definition of the core makes no mention of prices. However, if an exchange economy is sufficiently large, it is a remarkable fact that every core allocation is closely associated with a price vector that 'approximately decentralizes' it; in other words, every core allocation approximately satisfies the definition of Walrasian equilibrium, *without transfers*. This is an important strengthening of the Second Welfare Theorem. The notion of approximate decentralization depends to a considerable extent on the assumptions one is willing to make on the preferences and endowments of the individuals in the economy. (One version states that core allocations can be realized as *exact* Walrasian equilibrium with *small* income transfers.)

Core convergence has a number of implications, both normative and positive. The extent to which each of these implications is justified in a particular setting depends a great deal on the form of convergence, and thus on the assumptions one is willing to make on the economy. For an extensive survey focusing on the relationship between assumptions and the form of convergence, see Anderson (1992).

On the normative side, core convergence is a strong 'unbiasedness' property of Walrasian equilibrium, since it asserts that restricting attention to Walrasian allocations does not narrow the set of outcomes much beyond the narrowing that occurs in the core. Thus, Walrasian equilibrium has no hidden implications for the welfare of different groups, beyond whatever equity concerns one might have over the initial endowments. If one accepts the distribution of initial endowments as equitable, then any allocation that is far from Walrasian will not be in the core, and hence will treat some group of agents unfairly. On the positive side, if one accepts the core as a positive description of the allocations one is likely to see in practice in any economy, then core convergence tells us that the allocations we see will be nearly Walrasian.

However, the greatest significance of core convergence is as a test of the reasonableness of the price-taking assumption that is hidden in plain sight in the definition of Walrasian equilibrium. In real markets, we see prices used to equate supply and demand, but this does not guarantee Walrasian outcomes. Agents possessing market power may choose to demand quantities different from their price-taking demands at the prevailing price, thereby altering that price and leading to a non-Walrasian outcome. If the outcome is not at least approximately Walrasian, then the welfare theorems and the results on existence and generic determinacy of Walrasian allocations would have limited implications for real economies.

Core convergence and non-convergence allows us to identify situations in which price-taking is more or less reasonable. Core convergence implies that all trade takes place at almost a single price. An agent who tries to bargain cannot influence the prices much, so there is little incentive to be anything other than a price-taker. On the other hand, core non-convergence makes price-taking an implausible assumption.

Edgeworth (1881) doubted the positive significance of Walrasian equilibrium, and argued that the core, not the set of Walrasian equilibria, was the best positive description of the outcomes of a market mechanism. Moreover, while Edgeworth's name is closely associated with core convergence, and he did prove a core convergence theorem, he argued that in real economies, the presence of firms and syndicates which possess market power ensures that the core does *not* converge.

Edgeworth's argument about the effects of market power applies most strongly to the production side of the economy, where we do in fact see large firms, syndicates and labour unions. However, on the consumption side, the wealthiest individual in the world consumes a small part of the world's annual consumption. In exchange economies in which each consumer is small, core convergence holds. So core convergence provides a justification for the price-taking assumption on the consumption side, provided one views the world as an exchange economy in which the production decisions have been previously made by some exogenous process, outside the scope of the model, endowments include the income obtained by selling one's labour in the exogenous production process, and the only economic activity is trade and consumption of what has been produced.

The proof of the most basic core convergence theorem, which assumes very little about preferences and endowments, and establishes approximate decentralization in a relatively weak sense, is closely analogous to the proof of the Second Welfare Theorem. The approximately decentralizing price vector appears magically, as a consequence of the separating hyperplane theorem and the Shapley–Folkman theorem, which asserts that the sum of a large number of sets is approximately convex. Convexity of preferences plays no role. Indeed, the definition of the core, because it allows for individuals to be included or excluded from potential coalitions, introduces a non-convexity which is not present in the Second Welfare Theorem, and the Shapley–Folkman theorem controls that non-convexity, whether or not preferences themselves are convex.

The definitions and results just described verbally are presented more formally below.

Many people have made important contributions to the study of core convergence. A survey of these contributions is given in Anderson (1992), and a list of some of the more important contributions is included in the bibliography.

Now, We Turn to a More Formal Presentation

Definition 1 In an exchange economy with agents $i = 1, \ldots, I$ having strict preferences $\succ i$ and endowments $\omega_i \in \mathbf{R}^L_+$, a *coalition* is a set $S \subseteq \{1, \ldots, I\}$. An *exact allocation* is $x \in \left(\mathbf{R}^L_+\right)^I$ such that $\sum_{i=1}^I x_i = \sum_{i=1}^I \omega_i$. An exact allocation is *weakly Pareto optimal* if there is no other exact allocation x' satisfying $x'_i \succ {}_i x_i (i = 1, \ldots, I)$. A coalition S *blocks* or *improves on* an exact allocation x by x' if $\sum_{i \in S} x'_i = \sum_{i \in S} \omega_i$ and $\forall_{i \in S}\ x'_i \succ {}_i x_i$. The *core* is the set of all exact allocations which cannot be improved

on by any nonempty coalition. The *price simplex* is $\Delta = \left\{p \in \mathbf{R}_+^k : \sum_{\ell=1}^{L} p_\ell = 1\right\}$.

Theorem 2 *In an exchange economy, every core allocation is weakly Pareto optimal.*

Proof If x is not weakly Pareto optimal, then there exists x' such that $\sum_{i=1}^{I} x_i' = \sum_{i=1}^{I} x_i$, $x_i' \succ_i x_i$. Then $S = \{1, \ldots, I\}$ improves on x by x', so x is not in the core.

Theorem 3 (Strong First Welfare Theorem) *In an exchange economy, every Walrasian Equilibrium lies in the core.*

Proof Suppose (p^*, x^*) is a Walrasian Equilibrium. If x^* is not in the core, there exists $S \subseteq I$, $S \neq \varnothing$ and $x_i'(i \in S)$ such that $\sum_{i \in S} x_i' = \sum_{i \in S} \omega_i$ and $x_i' \succ_i x_i^*$ for each $i \in S$. Since x_i^* lies in i's demand set at the price p^*, $p^* \cdot x_i > p^* \omega_i$, so $p^* \cdot \sum_{i \in S} x_i' = \sum_{i \in S} p^* \cdot x_i > \sum_{i \in S} p^* \cdot \omega_i = p^* \cdot \sum_{i \in S} \omega_i$ but $\sum_{i \in S} \omega_i$, a contradiction. Therefore, x^* is in the core.♦

Theorem 4 (Core convergence, E. Dierker 1975, and Anderson 1978) Suppose we are given an exchange economy with L commodities, I agents and preferences $\succ_1, \ldots, \succ_I$ satisfying weak monotonicity (if $x \gg y$, then $x \succ_i y$) and the following free disposal condition: $x \gg y, y \succ_i z \Rightarrow x \succ_i z$. If x is in the core, then there exists $p \in \Delta$ such that

$$\frac{1}{I}\sum_{i=1}^{I} |p \cdot (x_i - \omega_i)| \leq \frac{2L}{I} \max\{||\omega_1||_\infty, \ldots ||\omega_I||_\infty\} \tag{1}$$

$$\frac{1}{I}\sum_{i=1}^{I} \left|\inf\{p \cdot (y - x_i) : y \succ_i x_i\right| \leq \frac{4L}{I} \max\{||\omega_1||_\infty, \ldots ||\omega_I||_\infty\} \tag{2}$$

where $||x||_\infty = \max\{|x_1|, \ldots, |x_L|\}$.

If there are many more agents than goods, and the endowments are not too large, the bounds on the right-hand sides of Eqs. (1) and (2) will be small. In that case, Eq. (1) says that trade occurs almost at the price p, and that each x_i is almost in the budget set, while Eq. (2) says that the price p almost supports $\succ_i$ at x_i, in the sense that everything preferred to x_i costs almost as much as x_i. Taken together, Eqs. (1) and (2) say that the pair (p, x) satisfies a slightly perturbed version of the *def* of Walrasian equilibrium. Indeed, if we knew the left sides of Eqs. (1) and (2) were *zero*, then $p \cdot (x_i - w_i) = 0$, so x_i lies in i's budget set, and $y \succ_i x_i \Rightarrow p \cdot y \geq p \cdot \omega_i$, so x would be a Walrasian quasi-equilibrium! (A pair (p^*, x^*) is said to be a Walrasian quasi-equilibrium if it satisfies the definition of a Walrasian equilibrium except that instead of requiring that x_i^* lie in i's demand set, we only require that x_i^* lie in i's quasi-demand set, that is $p^* \cdot x_i^* \leq p^* \cdot \omega_i$ and every $y \succ_i x_i^*$ satisfies $p^* \cdot y \leq p^* \cdot \omega_i$.)

Outline of Proof Follow the proof of the Second Welfare Theorem.

- Suppose x is in the core. Define $B_i = \{y - \omega_i : y \succ_i x_i\} \cup \{0\} = (\{y : y \succ_i x_i\} \cup \{\omega_i\}) - \omega_i$ and $B = \sum_{i=1}^{I} B_i$. The first term in the definition of B_i corresponds to members of a potential improving coalition; for accounting purposes, we assign members outside the coalition their endowments. Note that B_i is *not* convex, even if $\succ_i$ is a convex preference.

Claim If x is in the core, then $B \cap \boldsymbol{R}_{--}^L = \varnothing$. Suppose $z \in B \cap \boldsymbol{R}_{--}^L$. Then there exists $z_i = B_i$ such that $z = \sum_{i=1}^{I} z_i$. Let $S = \{i : z_i \neq 0\}$; since $z \ll 0$, $S \neq \varnothing$. For $i \in S$, let $x_i' = \omega_i + z_i - \frac{z}{|S|}$. Then $x_i' \gg \omega_i + z_i \succ_i x_i$ by the definition of B_i, $x_i' \succ_i x_i$ by free disposal, and $\sum_{i \in S} x_i' = \sum_{i \in S} \omega_i$, so S can improve on x by x', so x is not in the core.

- Let $v = -L(\max_{i=1,\&,I}||\omega_i||_\infty, \ldots, \max_{i=1,\&,I} ||\omega_i||_\infty$.

Claim $(\text{con } B) \cap \left(v + \boldsymbol{R}_{--}^L\right) = \varnothing$. If $z \in \text{con} B$, by the Shapley–Folkman theorem, and relabelling the agents, we may write

$$z = \sum_{i=1}^{I} z_i, \quad z_i \in \text{con} \quad B_i (i = 1, \ldots, I),$$
$$z_i \in B_i \quad (i \neq \{1, \ldots, L\})$$

Choose

$$\widehat{z_i} = \begin{cases} 0 & \text{if} \quad i = 1, \ldots, L \\ z_i & \text{if} \quad i = L+1, \ldots, I \end{cases}$$

Then $\sum_{i=1}^{I} \widehat{z_i} \in B$ so $\sum_{i=1}^{I} \widehat{z_i} \not\leq / < 0$. If $z \ll v$, then, $\sum_{i=1}^{I} \widehat{z_i} = \sum_{i=1}^{L} 0 + \sum_{i=L+1}^{I} z_i \leq \sum_{i=1}^{L} (\omega_i + z_i) + \sum_{i=L+1}^{I} z_i = \sum_{i=1}^{L} \omega_i \, \text{c} + \sum_{i=1}^{I} z_i = \sum_{i=1}^{L} \omega_i + z \ll \sum_{i=1}^{L} \omega_i + v \leq 0$, so $B \cap \mathbf{R}^L_{--} \neq \varnothing$ a contradiction which proves the claim.

- By the separating hyperplane theorem, there exists $p \neq 0$ such that $\sup p \cdot \left(v + \mathbf{R}^L_{--}\right) \leq \inf p \cdot (\text{con}\, B)$. If $p_\ell < 0$ for some ℓ, then $\sup p \cdot \left(v + \mathbf{R}^L_{--}\right) = +\infty$, while $\inf p \cdot (\text{con}\ B) \leq 0$, a contradiction, so $\text{p} = 0$ and we can normalize $p \in \Delta$. Then $\inf p \cdot B \geq \inf p \cdot (\text{con}\ B) \geq p \cdot v = -L \max \left\{ \|\omega_1\|_{\infty, \ldots,} \|\omega_I\|_{\infty} \right\}$.
- Adapt the remainder of the proof of the Second Welfare Theorem; this requires a few tricks.

See Also

- ▶ Arrow–Debreu Model of General Equilibrium
- ▶ Cores
- ▶ Edgeworth, Francis Ysidro (1845–1926)
- ▶ Existence of General Equilibrium
- ▶ General Equilibrium
- ▶ General Equilibrium (New Developments)

Bibliography

Anderson, R.M. 1978. An elementary core equivalence theorem. *Econometrica* 46: 1483–1487.

Anderson, R.M. 1981. Core theory with strongly convex preferences. *Econometrica* 49: 1457–1468.

Anderson, R.M. 1985. Strong core theorems with nonconvex preferences. *Econometrica* 53: 1283–1294.

Anderson, R.M. 1986. Core allocations and small income transfers. Working Paper No. 8621, Department of Economics, University of California at Berkeley.

Anderson, R.M. 1987. Gap-minimizing prices and quadratic core convergence. *Journal of Mathematical Economics* 16: 1–15. Correction, *Journal of Mathematical Economics* 20(1991): 599–601.

Anderson, R.M. 1992. The core in perfectly competitive economies. In *Handbook of game theory with economic applications*, vol. I, ed. R.J. Aumann and S. Hart. Amsterdam: North-Holland.

Aumann, R.J. 1964. Markets with a continuum of traders. *Econometrica* 32: 39–50.

Bewley, T.F. 1973. Edgeworth's conjecture. *Econometrica* 41: 425–454.

Brown, D.J., and A. Robinson. 1974. The cores of large standard exchange economies. *Journal of Economic Theory* 9: 245–254.

Brown, D.J., and A. Robinson. 1975. Nonstandard exchange economies. *Econometrica* 43: 41–55.

Debreu, G. 1975. The rate of convergence of the core of an economy. *Journal of Mathematical Economics* 2: 1–7.

Debreu, G., and H. Scarf. 1963. A limit theorem on the core of an economy. *International Economic Review* 4: 236–246.

Dierker, E. 1975. Gains and losses at core allocations. *Journal of Mathematical Economics* 2: 119–128.

Edgeworth, F.Y. 1881. *Mathematical psychics*. London: Kegan Paul.

Grodal, B. 1975. The rate of convergence of the core for a purely competitive sequence of economies. *Journal of Mathematical Economics* 2: 171–186.

Grodal, B. and W. Hildenbrand. 1974. Limit theorems for approximate cores. Working Paper IP-208, Center for Research in Management, University of California, Berkeley.

Hildenbrand, W. 1974. *Core and Equilibria of a large economy*. Princeton: Princeton University Press.

Kannai, Y. 1970. Continuity properties of the core of a market. *Econometrica* 38: 791–815.

Khan, M.A. 1974. Some equivalence theorems. *Review of Economic Studies* 41: 549–565.

Vind, K. 1964. Edgeworth allocations in an exchange economy with many traders. *International Economic Review* 5: 165–177.

Vind, K. 1965. A theorem on the core of an economy. *Review of Economic Studies* 32: 47–48.

Walras, L. 1874. *Eléments d'économie politique pure*. Lausanne: L. Corbaz.

Cores

Werner Hildenbrand

Abstract

The *core* of an economy consists of those states of the economy which no group of agents can 'improve upon'. The core is a rather theoretical

fundamental equilibrium concept. Indeed, the core provides a theoretical foundation of a more operational equilibrium concept, namely, the competitive equilibrium, which is a very different notion of equilibrium.

Keywords

Barter; Coalitions; Cobb–Douglas functions; Competitive equilibrium; Continuum of agents; Cooperative game theory; Cooperative game theory (core); Cores; Edgeworth, F.; Hildenbrand, W.; Initial endowments; Large economies; Limit economy; Limit theorems; Minkowski's separation theorem; Pareto efficiency; Preference relations; Replica economies; Type economies

JEL Classifications

D5

The *core* of an economy consists of those states of the economy which no group of agents can 'improve upon'. A group of agents can improve upon a state of the economy if, by using the means available to that group, each member can be made better off. Nothing is said in this definition of how a state in the core actually is reached. The actual process of economic transactions is not considered explicitly.

To keep the presentation as simple as possible, we shall consider only the core for exchange economies with an arbitrary number l of commodities, even though the core concept applies to more general situations.

Consider a finite set A of economic agents; each agent a in A is described by his *preference relation* $\precsim_a$ (defined on the positive orthant R^l_+) and his *initial endowments* $e_{a.}$ (a vector in R^l_+). The outcome of any exchange, that is to say, a state (x_a) of the exchange economy $\mathscr{E} = \{\precsim_a, e_a\}_{a \in A}$, is a *redistribution* of the total endowments, i.e.

$$\sum_{a \in A} x_a = \sum_{a \in A} e_a.$$

A *coalition* of agents, say $S \subset A$, can *improve upon* a redistribution (x_a), if that coalition S, by using the endowments available to it, can make each member of that coalition better off, that is to say, there is a redistribution, say $(y_a)_{a \in S}$, such that

$$y_a \succ_a x_a \text{ for every } a \in S \text{ and } \sum_{a \in S} y_a = \sum_{a \in S} e_a.$$

The set of redistributions for the exchange economy $\mathscr{E}$ that no coalition can improve upon is called the core of the economy $\mathscr{E}$, and is denoted by $C(\mathscr{E})$.

The core is a rather theoretical, however, fundamental equilibrium concept. Indeed, the core provides a theoretical foundation of a more operational equilibrium concept, the *competitive equilibrium* which, in fact, is a very different notion of equilibrium. The allocation process is organized through markets; there is a price for every commodity. All economic agents take the price system as given and make their decisions independently of each other. The equilibrium price system coordinates these independent decisions in such a way that all markets are simultaneously balanced.

More formally, an allocation $\left(x^*_a\right)$ for the exchange economy $\mathscr{E} = \{\precsim_a, e_a\}_{a \in A}$ is a *competitive equilibrium* (or a *Walras allocation)* if there exists a price vector $p^* \in R^l_+$ such that for every $a \in A, x^*_a \in \varphi_a(p^*)$ and

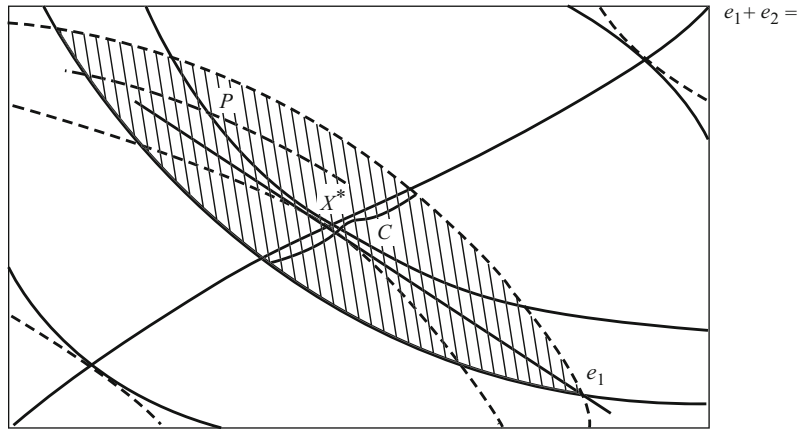
$$\sum_{a \in A} x^*_a = \sum_{a \in A} e_a.$$

Here $\varphi_a(p^*)$ or more explicitly, $\varphi(p^*, e_a, \precsim_a)$ denotes the demand of agent a with preferences $\precsim_a$ and endowment e_a, i.e. the set of most desired commodity vectors (with respect to $\precsim_a$) in the budget-set $\left\{x \in R^l_+ \mid p^* \cdot x \leq p^* \cdot e_a\right\}$.

The set of all competitive equilibria for the economy $\mathscr{E}$ is denoted by $W(\mathscr{E})$.

The core and the set of competitive equilibria for an economy with two agents and two commodities can be represented geometrically by the well-known Edgeworth–Box (see Fig. 1). The size of the box is determined by the total endowments $e_1 + e_2$. Every point P in the box represents

Cores, Fig. 1



a redistribution; the first agent receives $x_1 = P$ and the second receives $x_2 = (e_1 + e_2) - P$.

It is easy to show that for every exchange economy $\mathscr{E}$ a competitive equilibrium belongs to the core,

$$W(\mathscr{E}) \subset C(\mathscr{E}).$$

Thus, a state of the economy $\mathscr{E}$ which is decentralized by a price system cannot be improved upon by cooperation. This proposition strengthens a well-known result of Welfare Economics – every competitive equilibrium is Pareto-efficient.

The inclusion $W(\mathscr{E}) \subset C(\mathscr{E})$ is typically strict. Indeed, if the initial allocation of endowments is not Pareto-efficient, which is the typical case, then, if there are any allocations in the core at all, there are core-allocations which are not competitive equilibria.

This leads us to the *basic problem* in the theory of the core:

> For which kind of economies is the 'difference' between the core and the set of competitive equilibria small? Or in other words, under which circumstances do cooperative barter and competition through decentralized markets lead essentially to the same result?

Naturally, the answer depends on the way one measures the 'difference' between the two equilibrium concepts. However this is done one expects that the economy must have a large number of participants.

In answering the basic questions we try to be comprehensible (for example by avoiding the use of measure-theoretic concepts) but not comprehensive. Therefore, if we refer in the remainder of this entry to an economy $\mathscr{E} = \{\precsim_a, e_a\}_{a \in A}$ we shall always assume that preference relations are continuous, complete, transitive, monotone and strictly convex. The total endowments $\Sigma_{a \in A} e_a$ of an economy are always assumed to be strictly positive. We shall not repeat these assumptions. Furthermore, if we call an economy smooth, then we assume in addition that preferences are smooth (hence representable by sufficiently differentiable utility functions) and individual endowments are strictly positive.

These assumptions simplify the presentation tremendously. For generalizations we refer to the extensive literature.

We remark that under the above assumptions there always exists a competitive equilibrium, and hence, the core is not empty.

Large Economies

The simplest and most stringent measure of difference between the two equilibrium sets, $C(\mathscr{E})$, and $W(\mathscr{E})$, which we shall denote by $\delta(\mathscr{E})$, can be defined as follows.

Let $\delta(\mathscr{E})$ be the smallest number δ with the property: for every allocation $(x_a) \in C(\mathscr{E})$ there exists an allocation $\left(x_a^*\right) \in W(\mathscr{E})$ such that

$$|x_a - x_a^*| \leq \delta$$

for every agent a in the economy $\mathscr{E}$.

Thus, if $\delta(\mathscr{E})$ is small, then from every agent's view a core allocation is like a competitive equilibrium.

Unfortunately for this measure of difference, it is not true that $\delta(\mathscr{E})$ can be made arbitrarily small provided the number of agents in the economy $\mathscr{E}$ is sufficiently large (even if one restricts the agents' characteristics $(\precsim_a, e_a)$ to an a priori given finite set).

Consequently one considers also weaker measures for the 'difference' between the two equilibrium concepts $C(\mathscr{E})$ and $W(\mathscr{E})$ For example, define $\delta_1(\mathscr{E})$ and $\delta_2(\mathscr{E})$, respectively, as the smallest number δ with the property: for every $(x_a) \in C(\mathscr{E})$ there exists a price vector $p \in R^l_+$ such that

$$(\delta_1)|x_a - \varphi_a(p)| \leq \delta \text{ for every agent } a \text{ in } \mathscr{E}$$

or

$$(\delta_2)\frac{1}{\#A}\sum_{a \in A}|x_a - \varphi_a(p)| \leq \delta.$$

Clearly, the measures δ_1 and δ_2 are weaker than δ since the price vector p is not required to be an equilibrium price vector for the economy $\mathscr{E}$. The number $\delta_1(\mathscr{E})$ (and, *a fortiori*, $\delta_2(\mathscr{E})$) does not measure the distance between the sets $C(\mathscr{E})$ and $W(\mathscr{E})$. But the degree by which an allocation in the core can be decentralized via a price system. Obviously one has $\delta_2(\mathscr{E}) \leq \delta_1(\mathscr{E}) \leq \delta(\mathscr{E})$.

One can show that $\delta_2(\mathscr{E})$ becomes arbitrarily small for sufficiently large economies. More precisely,

Theorem 1 Let T be a finite set of agents' characteristics $(\precsim, e)$ and let b be a strictly positive vector in R^l. Then for every $\varepsilon > 0$ there exists an integer N such that for every economy $\mathscr{E} = \{\precsim_a, e_a\}_{a \in A}$ with $\#A \geq N$,

$$\frac{1}{\#A}\sum_{a \in A} e_a \geq b$$

and $(\succsim_a, e_a) \in T$ one has.

$$\delta_2(\mathscr{E}) \leq \varepsilon.$$

(The finite set T in Theorem 1 can be replaced by a compact set with respect to a suitably chosen topology: see Hildenbrand 1974.) We emphasize that this result does not imply that in large economies core-allocations are near to competitive equilibria. In fact, Theorem 1 does not hold if δ_2 is replaced by the measure of difference δ or even δ_1. Theorem 1 does imply, however, that for sufficiently large economies one can associate to every core-allocation a price vector which 'approximately decentralizes' the core-allocation. Some readers might consider this conclusion as a perfectly satisfactory answer to our basic problem. If one holds this view, then the rest of the paper is a superfluous intellectual pastime. We would like to emphasize, however, that the meaning of 'approximate decentralization' is not very strong. First, the demand $\varphi_a(p)$ is not necessarily near to x_a for every agent a in the economy; only the mean deviation

$$\frac{1}{\#A}\sum_{a \in A}|x_a - \varphi_a(p)|$$

becomes small. Second, total demand is not equal to total supply; only the mean excess demand

$$\frac{1}{\#A}\sum_{a \in A}[\varphi_a(p) - e_a]$$

becomes small.

There are alternative proofs in the literature, e.g. Bewley (1973), Hildenbrand (1974), Anderson (1981) or Hildenbrand (1982). These proofs are based either on a result by Vind (1965) or Anderson (1978).

Sharper conclusions than the one in Theorem 1 will be stated in the following sections. There we consider a sequence $(\mathscr{E}_n)_{n=1,\ldots}$ of economies and then study the asymptotic behaviour of $\delta(\mathscr{E}_n)$.

Before we present these limit theorems we should mention another approach of analysing the inclusion $W(\mathscr{E}) \subset C(\mathscr{E})$. Instead of analysing the asymptotic behaviour of the difference $\delta(\mathscr{E}_n)$ for a sequence of finite economies one can define a large economy where every agent has strictly no

C

influence on collective actions. This leads to a *measure space without atoms* of economic agents (also called a *continuum of agents).* For such economies the two equilibrium concepts coincide. See Aumann (1964).

Replica Economies

Let $\mathscr{E}; = \{\precsim_i, e_i\}$ be an exchange economy with m agents. For every integer n we define the n-fold *replica economy* $\mathscr{E}_n$ of $\mathscr{E}$ as an economy with $n{\cdot}m$ agents; there are exactly n agents with characteristics $(\precsim_i, e_i)$ for every $i = 1,\ldots, m$.

More formally,

$$\mathscr{E}_n = \left\{\precsim_{(i,j)}, e_{(i,j)}\right\} \begin{matrix} 1 \le i \le m \\ 1 \le i \le n \end{matrix}$$

where $\precsim_{(i,j)} = \precsim_i$ and $e_{(i,j)} = e_i$, $1 \le i \le m$ and $1 \le j \le n$. Thus, an agent a in the economy $\mathscr{E}_n$ is denoted by a double index $a = (i,j)$. We shall refer to agent (i,j) sometimes as the jth agent of type i.

Replica economies were first analysed by F. Edgeworth (1881) who proved a limit theorem for such sequences in the case of two commodities and two types of agents. A precise formulation of Edgeworth's analysis and the generalization to an arbitrary finit/e number of commodities and types of agents is due to Debreu and Scarf (1963).

Here is the basic result for replica economies.

Theorem 2 For every sequence $(\mathscr{E}_n)$ of replica economies the difference between the core and the set of competitive equilibria tends to zero, i.e.,

$$\lim_{n\to\infty} \delta(\mathscr{E}_n) = 0.$$

Furthermore, if $\mathscr{E}$ is a smooth and regular economy then $\delta(\mathscr{E}_n)$ converges to zero at least as fast as the inverse of the number of participants, i.e., there is a constant K such that

$$\delta(\mathscr{E}_n) \le \frac{K}{n}.$$

The proof of this rmkably neat result is based on the fact that a core–allocation (x_{ij}) assigns to every agent of the same type the same commodity bundle, i.e., $x_{ij} = x_{ik}\,\mathscr{E}$, This 'equal treatment' property simplifies the analysis of $\delta(\mathscr{E}_n)$ tremendously. Indeed, an allocation (x_{ij}) in $C(\mathscr{E}_n)$, which can be considered as a vector in $R^{l\cdot m\cdot n}$, is completely described by the commodity bundle of one agent in each type, thus by a vector $(x_{11}, x_{21},\ldots, x_{m1})$ in $R^{l\cdot m}$, a space whose dimension is independent of n.

Thus, let

$$C_n = \left\{(x_{11}, x_{21}, \ldots, x_{m1}) \in R^{l\cdot m} \middle| \left(x_{ij}\right) \in C(\mathscr{E}_n)\right\}.$$

One easily shows that $C_{n+1} \subset C_n$. It is not hard to see that Theorem 1 follows if

$$\cap_{n=1}^{\infty} C_n = W(\mathscr{E}_1).$$

But this is the well-known theorem of Debreu and Scarf (1963). The essential arguments in the proof go as follows. Let $(x_1, \ldots, x_m) \in \cap_{n=1}^{\infty} C_n$. One has to show that there is a price vector p^* such that $x \succ_i x_i$ implies $p^* \cdot x > p^* \cdot e_i$ For this it suffices to show that there is a p^* such that

$$p^* \cdot z \ge 0 \text{ for every } z \in \cup_{i=1}^{m} \left(\left\{x \in R^l_+ \mid x \succ_i x_i\right\}\right) - e_i) = Z,$$

i.e., there is a hyperplane (whose normal is p^*) which supports the set z. One shows that the assumption $(x_1, \ldots, x_m) \in \cap_{n=1}^{\infty} C_n$ implies that 0 does not belong to the convex hull of z. Minkowski's Separation Theorem for convex sets then implies the existence of the desired vector p^*.

The second part of the conclusion of Theorem 2 is due to Debreu (1975).

Type Economies

The limit theorem on the core for replica economies is not fully satisfactory since replication is a very rigid way of enlarging an economy. The conclusion '$\delta(\mathscr{E}_n) \to 0$' in Theorem 2, to be of general relevance, should be robust to small deviations from the strict replication procedure.

Consider a sequence $(\mathscr{E}_n)$ of economies where the characteristics of every agent belong to a given finite set of types $T = \{(\precsim_1, e_1), \ldots, (\precsim_m, e_m)\}$. We do not consider this as a restrictive assumption (considered as an approximation, one can always group agents' characteristics into a finite set of types). Let the economy $\mathscr{E}_n$ have N_n agents; $N_n(1)$-agents of the first type, $N_n(i)$ agents of type i. Of course the idea is that N_n tends to ∞ with increasing n. Consider the fraction $\nu_n(i)$ of agents in the economy $\mathscr{E}_n$ which are of type i, i.e.,

$$\nu_n(i) = \frac{N_n(i)}{N_n}.$$

The sequence $(\mathscr{E}_n)$ is a replica sequence of an economy $\mathscr{E}$ (not necessarily of $\mathscr{E}_1$) if and only if the fractions $\nu_n(i)$ are all independent of n. It is this rigidity which we want to weaken now.

A sequence $(\mathscr{E}_n)$ of economies with characteristics in a finite set T is called a *sequence of type economies* (over T) if

(i) the number N_n of agents in $(\mathscr{E}_n)$ tends to infinity and

(ii) $\nu_n(i) = \frac{N_n(i)}{N_n} \underset{(n\to\infty)}{\to} \nu(i) > 0.$

EX (random sampling of agents' characteristics):

Let π be a probability distribution over the finite set T. Define the economy $\mathscr{E}_n$ as a random sample of size n from this distribution $\pi(\cdot)$. The law of large numbers them implies property (ii): $\nu_n(i) \to \pi(i)$.

The step from replica economies to type economies – as small as it might appear to the reader – is conceptually very important. Yet with this 'small' generalization the analysis of the limit behaviour of $\delta(\mathscr{E}_n)$ or $\delta_1(\mathscr{E}_n)$ is made more difficult. Even worse, it is no longer true that for *every* sequence $(\mathscr{E}_n)$ of type economies one obtains $\delta(\mathscr{E}_n) \to 0$ – even if the preferences of all types are assumed to be very nice, say smooth. There are some 'exceptional cases' where the conclusion $\delta(\mathscr{E}_n) \to 0$ does not hold. But these are 'exceptional' cases and the whole difficulty in the remainder of this section is to explain in which precise sense these cases are 'exceptional' and can therefore be ignored. We shall first exhibit the 'cases' where the conclusion fails to hold. Then we shall show that these cases are exceptional.

We denote by $\Pi(\mathscr{E})$ the set of normalized *equilibrium price vectors* for the economy $\mathscr{E} = \{\precsim_a, e_a\}_{a\in A}$ Thus, for $p^* \in \Pi(\mathscr{E})$ the excess demand is zero, i.e.,

$$\sum_{a\in A}[\varphi_a(p^*) - e_a] = 0.$$

To every sequence $(\mathscr{E}_n)$ of type economies we associate a '*limit economy*' $\mathscr{E}_\infty$. This economy has an 'indefinitely large' number of agents of every type; the fraction of agents of type i is given by $\nu(i)$. The mean (per capita) excess demand of that limit economy $\mathscr{E}_\infty$ is defined by

$$z_\nu(p) = \sum_{i=1}^{m} \nu(i)[\varphi(p, e_i, \precsim_i) - e_i].$$

An equilibrium price vector p^* of the limit economy $\mathscr{E}_\infty$ is defined by $z_\nu(p^*) = 0$. Let $\Pi(\nu)$ denote the set of normalized equilibrium price vectors for $\mathscr{E}_\infty$. Obviously for a replica sequence $(\mathscr{E}_n)$ we have $\Pi(\mathscr{E}_n) = \Pi(\nu)$ for all n. However, for a sequence of type economies the set $\Pi(\mathscr{E}_n)$ of equilibrium prices of the economy $\mathscr{E}_n$ depends on n, and it might happen that the set $\Pi(\nu)$ is not similar to $\Pi(\mathscr{E}_n)$ even for arbitrarily large n. To fix ideas, it might happen that $\Pi(\mathscr{E}_n) = \{p_n\}$ and $\Pi(\nu)$ contains not only $p = \lim p_n$ but also another equilibrium price vector. Such a situation has to be excluded.

We call a sequence of type economies *sleek* if $\Pi(\mathscr{E}_n)$ converges (in the Hausdorff-distance) to $\Pi(\nu)$.

It is known (Hildenbrand 1974) that the sequence $(\Pi(\mathscr{E}_n))$ converges to $\Pi(\nu)$ if $\Pi(\nu)$ is a singleton (i.e., the limit economy has a unique equilibrium) or, in general, if (and only if) for every open set O in R^ℓ with $O \cap \Pi(\nu) \neq \emptyset$ it follows that $O \cap \mathscr{E}\,\Pi(\mathscr{E}_n) \neq \emptyset$ for all n sufficiently large.

We now have exhibited the cases where a limit theorem on the core holds true.

Theorem 3 For every sleek sequence $(\mathscr{E}_n)$ of type economies

$$\lim_{n\to\infty} \delta(\mathscr{E}_n) = 0$$

Unfortunately there seems to be no short and easy proof. The main difficulty arises from the fact that for allocations in the core of a type economy the 'equal treatment' property, which made the replica case so manageable is no longer true. For a proof see Hildenbrand and Kirman (1976) or Hildenbrand (1982) and the references given there. The main step in the proof is based on a result of Bewley (1973).

It remains to show that non-sleek sequences of type economies are 'exceptional cases'.

The strongest form of 'exceptional' is, of course, 'never'. We mentioned already that a sequence $(\mathscr{E}_n)$ is sleek if its limit economy has a unique equilibrium. Unfortunately, however, only under very restrictive assumptions on the set T of agents' characteristics does uniqueness prevail; for example,

(1) if every preference relation leads to a demand function which satisfies gross-substitution (Cobb–Douglas utility functions are typical exs),
(2) if every preference relation is homothetic and the endowment vectors $e_i (i = 1,\ldots, m)$ are collinear.

Since there is no reasonable justification for restricting the set T to such special types of agents we have to formulate a model in which we allow non-sleek sequences to occur provided, of course, this can be shown to be 'exceptional cases'. Let S^{m-1} denote the open simplex in R^m, i.e.

$$S^{m-1} = \left\{ x \in R^m \mid x_i > 0, \sum_{i=1}^{m} x_i = 1 \right\}.$$

The limit distribution $\nu(i)$ of a sequence of type economies with m types is a point in S^{m-1}.

A closed subset C in S^{m-1} which has (m – 1 dimensional Lebesgue) measure zero is called *negligible*. Thus, if a distribution ν is not in C then a sufficiently small change will not lead to C. Furthermore, given any arbitrary small positive number $\mathscr{E}$ one can find a countable collection of balls in S^{m-1} such that their union covers C, and that the sum of the diameters of these balls is smaller than $\mathscr{E}$. Thus, in particular, if $\nu \in C$ then one can approximate ν by points which do not belong to C. Clearly, a negligible set is a small set in S^{m-1}.

Theorem 4 Given a finite set T of m smooth types of agents, there exists a negligible subset C in S^{m-1} and a constant K such that for every sequence $(\mathscr{E}_n)$ of type economies over T whose limit distribution ν does not belong to C one has $\delta(\mathscr{E}_n) \leq K/\#A_n$, thus in particular, $\lim_{n\to\infty} \delta(\mathscr{E}_n) = 0$.

The convergence of $\delta(\mathscr{E}_n)$ follows from Theorem 3 and Theorems 5.4.3 and 5.8.15 in Mas–Colell (1985). For the rate of convergence see Grodal (1975).

See Also

- Edgeworth, Francis Ysidro (1845–1926)
- Existence of General Equilibrium

Bibliography

There is an extensive literature on limit theorems on the core which contains important generalizations of the results given here. For a general reference we refer to Hildenbrand (1974) or (1982), Mas-Colell (1985), Anderson (1981) and the references given there.

Anderson, R.M. 1978. An elementary core equivalence theorem. *Econometrica* 46: 1483–1487.

Anderson, R.M. 1981. Core theory with strongly convex preferences. *Econometrica* 49: 1457–1468.

Aumann, R.J. 1964. Markets with a continuum of traders. *Econometrica* 32: 39–50.

Bewley, T.F. 1973. Edgeworth's conjecture. *Econometrica* 41: 425–454.

Debreu, G. 1975. The rate of convergence of the core of an economy. *Journal of Mathematical Economics* 2: 1–8.

Debreu, G., and H. Scarf. 1963. A limit theorem on the core of an economy. *International Economic Review* 4: 235–246.

Edgeworth, F.Y. 1881. *Mathematical psychics*. London: Kegan Paul.

Grodal, B. 1975. The rate of convergence of the core for a purely competitive sequence of economies. *Journal of Mathematical Economics* 2: 171–186.

Hildenbrand, W. 1974. *Core and equilibria of a large economy*. Princeton: Princeton University Press.
Hildenbrand, W. 1982. Core of an economy. In *Handbook of mathematical economics*, ed. K.J. Arrow and M.D. Intriligator, vol. 2. Amsterdam: North-Holland.
Hildenbrand, W., and A.P. Kirman. 1976. *Introduction to equilibrium analysis*. Amsterdam: North-Holland.
Mas-Colell, A. 1985. *The theory of general economic equilibrium, a differentiable approach*. Cambridge: Cambridge University Press.
Vind, K. 1965. A theorem on the core of an economy. *Review of Economic Studies* 32: 47–48.

Corn Laws

B. Hilton

The British Corn Laws were parliamentary statutes which attempted to regulate the trade in corn (mainly wheat, barley, rye, and oats) for the benefit of producers during periods of plenty and low prices. Legislation to prohibit or discourage importation can be traced back to the 15th century, though it only became effective with an Act of 1663, while bounties to encourage export date from the 14th century and became more systematic after an Act of 1689. However, no economic tracts or pamphlets seem to have been devoted exclusively to this subject before 1750 (Barnes 1930, p. 16) and it was only in the 19th century that such legislation became controversial, mainly because the growth of population and especially of towns fuelled the concern about food supply that had been provoked by the scarcity of 1795 and by Malthus's *Essay* of 1798. In particular the 1815 Corn Law, which aimed to encourage domestic production by prohibiting importation until home prices had reached a certain level (80 shillings per quarter in the case of wheat), was the object of violent abuse both from radicals representing the interests of the consumer and also from middle-class manufacturers and exporters. In practice the 1815 Law satisfied no one. The sudden switch from total prohibition to total freedom of import at a particular price was destabilizing and failed to safeguard supply, since by the time (usually October or November) that prices reached the specified level, signalling a scarcity, the Baltic Sea was likely to have frozen over making cheap foreign imports unavailable for the remainder of the season. To meet these problems a sliding scale of duties was introduced in 1828, modified downwards in 1842, and finally abandoned seven years later after a major political crisis in 1846 had brought down Sir Robert Peel's second government and fundamentally divided the Conservative Party. The repeal of the Corn Laws was considered to mark the final triumph of free trade theories in Britain and quickly acquired symbolic importance, though with the drying up of European wheat supplies from the mid-1830s the Corn Laws had ceased to make very much practical difference to the trade in grain (Fairlie 1965). Repeal did not (as was widely expected) lead to reductions in the price of wheat, though it did reduce fluctuations in the amounts annually imported.

An assessment of the place of the Corn Laws in political economy must depend on what is understood by 'political economy'. In the first half of the 19th century the term was often used in a vulgar sense, and often abusively, to denote the prescriptions of those members of the middle classes who were leading the attack on 'old corruption', as the monopoly of power and privilege by the landed elite was commonly termed. At this level the assault on protection generally, and on agricultural protection in particular, was the sharpest weapon in the political economist's armoury, and for many the work of the Anti-Corn Law League under Cobden and Bright, and the promptings of the *Economist* newspaper, marked 'the high tide of *laissez-faire*'. If, on the other hand, political economy is taken to denote a body of formal economic thought, the Corn Laws must be accorded considerably less significance. Smith had exempted food supply and defence from the areas of public life to which the maxim of free trade should be made to apply, and most of his 19th-century successors were similarly well disposed to the Laws. Malthus consistently defended them while protesting that in all other matters he was a friend to free trade; Senior, who of all economists was the one most engaged in

advising on public policy, and whose political views inclined him against the agricultural lobby, in fact had little to say on the subject. Even J.S. Mill, while welcoming Corn Law repeal, consistently played down the beneficial effects which were likely to flow from it (Blaug 1968, pp. 192, 215). Probably only Ricardo placed hostility to the Corn Laws at the centre of his system. His corn model postulated that without access to cheap supplies of foreign grain, population pressure would either force domestic farmers onto more marginal land or else would compel them to cultivate the old land more intensively. In either case prices would rise, money wages would rise, profits would decline, and the economy would move towards a stationary state. Similar arguments were made by Torrens in his early work, though later he significantly qualified his opposition to the Corn Laws, and also by McCulloch, though he was always more concerned that the Laws caused excessive price fluctuations, and he eventually came to reject most aspects of the Ricardian corn model (O'Brien 1970, pp. 378–95). Indeed it has been argued that even for Ricardo the corn model was essentially an abstraction, and that his real animus against the Laws was based on their contribution to price instability and the shelter which they afforded to inefficient producers (Hollander 1979, pp. 605, 629–37, 647).

Moreover, the reduction and then finally the repeal of the Corn Laws seems to have owed little to economic doctrine. As early as 1821 Lord Liverpool's government envisaged the gradual dismantling of a system which it had always regarded as designed largely to ease the transition from a prolonged state of warfare and de facto protection (1793–1815) to a state of peace. Its main concern was with the supply situation, especially the unreliability of Ireland as a granary and increasing dependence on farmers in northern Europe in lean seasons. A subsidiary but significant factor was the return to the gold standard in 1821. It was now thought desirable to render the corn trade as regular (i.e. as little weather-related) as possible, in order to minimize fluctuations in bullion outflows (Hilton 1977, pp. 98–126). Both these factors operated on Peel in the prelude to repeal in 1846. The Irish Famine of 1845–9 emphasized the precariousness of the situation with regard to domestic supply (though it served to confirm and excuse the policy of repeal rather than initiating it), while the Bank Charter Act of 1844 rendered the money supply more than ever sensitive to movements of bullion. Political factors also obtruded in the decision to repeal: especially, the Anti-Corn Law League's activities threatened a whig victory in the next general election unless something was done to undermine its existence (Prest 1977, pp. 72–102). Undoubtedly there was also a mythical element in the campaign, the Corn Laws having become a symbol rather than a real guarantee of landed monopoly (Kemp 1961–2). There was, however, one theoretical argument which may have counted in the decision. This was the adoption of a market theory of wages in preference to the subsistence theory derived from Ricardo and the labour theory of value. The market theory was hardly novel in the 1840s, having been espoused by Malthus, but it was not given public prominence until Cobden bruited it emphatically in his campaign for repeal. Thus Peel ascribed his change of heart on agricultural protection in part at least to a discovery that 'the wages of labour do not vary with the price of grain'. He seems also to have been moved by Cobden's claims that in free trade conditions agriculturists would find it easier to capitalize and to engage in the 'high farming' that would be their competitive salvation. Such confidence seemed to be justified by the 'Golden Age of English Farming' which succeeded repeal, the real challenge to agriculture not occurring until the appearance of imports from the New World in the last quarter of the 19th century. By then, however, the Corn Laws had taken their place with king Richard III and the Inquisition among the 'bad things' of history, so much so that the cry of 'cheap loaf' was to prove politically irresistible, and to impair all early 20th-century attempts at tariff reform and imperial preference.

See Also

- ▶ Cobden, Richard (1804–1865)
- ▶ Compensation Principle
- ▶ Manchester School

Bibliography

Barnes, D.G. 1938. *A history of the English Corn Laws from 1660–1846*. London: Routledge.
Blaug, M. 1968. *Economic theory in retrospect*, 2nd ed. London: Heinemann.
Fairlie, S. 1965. The nineteenth-century corn law reconsidered. *Economic History Review* 18-(December): 562–575.
Hilton, B. 1977. *Corn, cash, commerce. The economic policies of the Tory governments 1815–1830*. Oxford: Oxford University Press.
Hollander, S. 1970. *The economics of David Ricardo*. London: Heinemann.
Kemp, B. 1961–2. Reflections on the repeal of the corn laws. *Victorian Studies* 5(3), March, 189–204.
O'Brien, D.P. 1978. *J.R. McCulloch. A study in classical economics*. London: George Allen & Unwin.
Prest, J. 1977. *Politics in the age of Cobden*. London: Macmillan.

Corn Laws, Free Trade and Protectionism

John Nye

Abstract

In the 1840s, Britain repealed the export restrictions and import duties on wheat known as the Corn Laws. But the traditional story of British free trade was complicated by an unwillingness to eliminate the most binding tariffs on wine and other consumables. In contrast, Britain's avowedly protectionist rival France had a more liberal trade policy than did Britain for most of the 19th century. Only with the 1860 Anglo–French Treaty of Commerce did Britain and France both move to uniformly low tariffs on goods and services, ushering in a period of genuinely free trade throughout Europe.

Keywords

Anti-Corn Law League; Chevalier, M.; Cobden, R.; Comparative advantage; Corn Laws; Excise taxes; Free trade; Income tax; Industrial Revolution; Infant-industry protection; Mercantilism; Mun, T.; Protection; Ricardo, D.; Scottish Enlightenment; Smith, A.; Specie; Tariffs; Trade deficit; World Trade Organization

JEL Classifications

N4

The Corn Laws were the parliamentary statutes that regulated the import and export of grain for the benefit of British producers in the early 19th century. Though these laws derived from legislation in the period 1804–15, they were but the extension or modification of a system that had been introduced in 1773 to prohibit exports of wheat when prices rose above a given level and that limited imports through a variety of duties based on a sliding scale. The goal of these laws was ostensibly the desire to stabilize the price of grain, which had been a regular goal of parliament since the late 17th century.

The debates about the abolition of the Corn Laws in the early to mid-1800s hold a special place in the economic history of Great Britain on account of their central role in shifting commercial policy to nearly free trade. Because of Britain's dominance of industrial trade in the 19th century and the leadership she exerted in international commerce, the struggles over the Corn Laws have been seen as emblematic of all debates about the advisability of free trade or protectionism. Despite the symbolic importance of these events, it is easy to overlook the facts that Britain after the repeal of the Corn Laws did not immediately move to perfectly free trade and that the political struggle over their abolition had at

least as much to do with domestic concerns over the importance of agriculture in a modern economy as with ideological questions about the advisability of free trade.

Mercantilism and the Rise of British Liberalism

The regulation or promotion of international trade has been perhaps the oldest policy issue in the political economy of international relations.

It is a common belief that trade is a primary source of a nation's wealth. But this has often been misunderstood to mean that exports enhance wealth while imports detract from it. This view, a central component of what is called mercantilism, stems from the mistaken belief that the benefits of trade flow only one way. One view was that a nation's wealth derived from the quantity of specie or gold and silver coin in the country. Therefore, exports contributed to this while imports detracted from it.

Some of this reasoning was theoretical, but more commonly mercantile theory was simply the evolution of a set of policies deriving from the fiscal needs of the newly emerging nation-states. Unsurprisingly, many states viewed the success of the state as synonymous with the success of the nation itself. Revenue was essential to the maintenance of the large armies that were a prerequisite for the nation-state. So trade was viewed as an essentially zero-sum game with both losers and winners. Moreover, this concern about revenue often translated into a concern for specie. Whereas modern economics treats specie as virtually irrelevant to the supply of money, contemporaries viewed coin itself as a necessary prerequisite of sound financial policy. Hence trade surpluses were preferred because they brought more precious metals in than they took out of the kingdom.

One of the earliest theoretical discussions of this view comes from Thomas Mun, who wrote 'The ordinary means therefore to increase our wealth and treasure is by foreign trade, wherein we must ever observe this rule; to sell more to strangers yearly than we consume of theirs in value' (1664, p. 11).

Adam Smith, the founder of modern economics, was the most prominent critic of this view. Starting from the observation that voluntary trade was mutually beneficial, and noting that the wealth of a nation's inhabitants, not its quantity of coin, made for true wealth, Smith argued in the *Wealth of Nations* against what he labelled the 'mercantile system'. He articulated the virtues of free and open trade, both in international and in home commerce. Indeed, the term 'free trade' was employed throughout the 18th and 19th centuries to refer to unregulated domestic trade as least as often as it referred to the free flow of goods from abroad.

These ideas were later modelled more systematically by the English economist David Ricardo, who formalized the analysis and showed that nations could maximize their welfare by specializing in the production of goods with the lowest opportunity cost and trading with other nations. This is the central idea behind the law of comparative advantage, usually attributed to David Ricardo, and developed more thoroughly by Paul Samuelson and others in the 20th century. Most important for this claim was the idea that a nation did not even have to be the 'best' producer of any product for there to be gains from trade. A nation that was more productive that another in all industries would still do better by specializing in some areas and trading for the other goods with another country. Thus, any claim that a nation could not benefit if it had no comparative advantage would be false. Every nation has a comparative advantage in producing some product, even if it has an absolute advantage in none.

Smith's ideas and those of his successors provided the philosophical basis for the classical liberal movements of the late 18th and early 19th centuries. By the early 1800s, the idea of a limited state that minimized regulation and promoted welfare through the encouragement of open trade at home and abroad had emerged as an important ideological view, promoted by prominent intellectuals and supported by an influential subset of the British political class. Nonetheless, the strong interest in the liberal ideas derived from the Scottish Enlightenment persuaded states not to fully adopt a policy of free trade. This was often not so

much the result of any ideological predisposition as a response to the state's desire for greater revenue. Taxing trade – both at home and from abroad – was one of the most common means of generating the income that supported the expanding bureaucracy of the modern state. Furthermore, special interests often worked to distort policy to favour of specific producers or economic sectors.

Since the late 17th century, Britain had been especially dependent on customs and excise taxes of various sorts. The rise of British liberalism had come in the same century (the 18th) that had seen the British state grow to an unprecedented size. Growth of government revenue had vastly outstripped the rate of overall economic growth and served to fund a professional bureaucracy at home and an expanding imperialist policy abroad. This enabled the British to either defeat or stalemate their traditional rival, France, in a series of military struggles that extended from the late 1600s to the era of Napoleon a century later. Moreover, this expansion of the central government came with little change in the revenues from land, the traditional source of income. Most of the gains came from steep increases in revenue from trade; and rising excises were some of the abuses cited by the American colonists as the basis for the independence movement.

However, changes in the landscape of the British economy – most notably the urban and industrial expansion that began in the late 1700s and is known as the Industrial Revolution – made Britain the premier industrial producer of the early 19th century and put pressure on the government to transform legislation that had kept agricultural prices high and had limited imports for the benefit of the farmers who were an increasingly small share of the economy.

The 19th-Century Corn Law Repeal: Free Trade Rhetoric vs. Protectionist Reality

The interests of industrial producers who felt that workers would be better served by cheap bread and the ideas of liberal elites saw concrete expression in the creation of the Anti-Corn Law League beginning in the 1830s. Statesmen such as Richard Cobden explicitly saw the movement as the first step in an attempt to push the British government to adopt a general policy of free trade.

However, it is not clear that theoretical ideas played a large role in the actual dismantling of the Corn Laws. Furthermore, Smith had always held up the staple industries and national defence as areas that might be exceptions to the doctrine of pure laissez-faire. However, the end of the Napoleonic Wars in 1815 removed the basis for wartime support of the Corn Laws and pushed the government to consider modifying or abolishing the restrictions in a transition to a peacetime economy.

As early as 1821 the government of Lord Liverpool had begun to consider reforming a system that it regarded as temporary and motivated by a desire to secure stable prices during wartime with a mix of regulation and protection. The Corn Laws did not seem to be fulfilling that function and, in the absence of war, their maintenance seemed unnecessary for the public good. Of course, the farm interests that gained from these rules would have fought for the continuation of these protections. Nonetheless, the increased voting power of urban workers empowered by the 1832 Reform Act reinforced Prime Minister Peel's conviction that support for industry was vital to the future development of Britain and led him to push for the abolition of all Corn Laws in the 1840s. The onset of the Irish potato famine in 1845 gave a special impetus to the desire to promote lower prices for basic staples and allowed Peel to push for the full abolition of the Corn Laws in 1846.

This legislation repealing the Corn Laws is often cited as the pivotal moment in the rise of free trade in Britain and in Europe because it was followed over the next decade with the reduction or removal of duties on hundreds of imports in Britain – hence the claim that henceforth Britain moved swiftly to full free trade. However, this accomplishment has been somewhat exaggerated in conventional history. Partly because of the need for continued revenue and partly because of pressure from special interests, a few large and important tariffs on coffee, tea, wine, spirits, sugar and

tobacco continued up to the 1860s, tariffs which had a disproportionate impact on the trade of Britain.

The 1860 Anglo-French Trade Treaty and the True Coming of Free Trade

The wine and spirit tariffs were especially important and had been mentioned prominently in Smith's criticism of the mercantile system in the *Wealth of Nations*. These tariffs had arisen from Britain's desire to punish her rival France and had developed as a means of protecting domestic beverage interests such as beer and gin at home, and colonial imports such as rum. Lacking an equivalent slogan to that of the cry for 'cheap bread', there was no great movement to reform these substantial duties.

Consequently, despite the British reputation as the leading free trader in the 19th century, Britain in fact had higher average tariffs than the more openly interventionist nation of France for the first three quarters of that century. The burden on the working classes from the combination of high tariffs on imported wine and liquor and the regulation and taxation of domestic production meant that consumption of alcohol was repressed throughout the 18th and early 19th centuries, despite all the income gains during the Industrial Revolution. Where basic alcoholic beverages had been seen as a necessary staple in the 17th century, they were more likely to be treated as luxuries in the 19th.

Full reform had to wait until 1860, when Britain and France concluded the Anglo–French Treaty of Commerce. This landmark treaty can be said to have truly ushered in the age of free trade in Europe. Brokered by Cobden in Britain and Michel Chevalier in France, the treaty had come after many years of negotiation. Early overtures to the French to sign such a treaty had been rebuffed in the 1840s because Britain had been unwilling to compromise their duties on wine – which had been the category of greatest concern to the French. However, changes in British fiscal structure arising from the imposition of an income tax in the 1850s made it easier for the British government to contemplate tariff cuts that might have compromised the budget in the short run. (British Liberals believed that given enough time, lower rates on imported wine would be offset by increased trade, a belief that proved accurate.) Moreover, the political considerations that led to wine duties being designed from the early 1700s to favour the products of friendly nations such as Portugal and Spain over that of France grew less important in the decades of peace following the defeat of Napoleon Bonaparte in 1815.

Thus, it became possible to conclude a treaty in 1860 in which Britain lowered and modified all its wine and spirit tariffs to remove any anti-French bias and caused France to lower tariffs and remove all prohibitions on goods – primarily textiles – imported from Britain. The 1860 Treaty was also significant for being a Most Favoured Nation agreement in which any subsequent treaties with third countries negotiated by either party would cause concessions to be applied equally to the original signatories. Concern by other Western nations that they would be left out of a trading arrangement between the two leading European powers led to almost the whole of Europe concluding equivalent treaties with either Britain or France over the next decade. By the 1870s virtually the whole of Europe was an extremely open trading area with free movement of goods, capital, and labour that in some ways has never been matched even by today's European Union. And by the end of the 19th century Britain could be said to have genuinely become a free trader with few or modest tariffs on most items, and possibly the lowest average tariffs in all Europe.

It is also interesting to note that Britain provides something of a counterexample to the tendency of modern-day protectionists to fret about the trade balance. Britain was the undoubted leader in world trade throughout the 19th century yet she also ran a merchandise trade deficit for virtually the whole of that period up to the First World War.

The one major counter-example to the tendency in the West to move towards freer international commerce had been the United States. Whereas Europe was busy lowering or abolishing tariffs and trade restrictions after 1860, the USA raised tariffs substantially from the 1860s

onwards. Tariffs were the major source of revenue for the federal government before the constitutional amendment that permitted an income tax. Furthermore, the civil war gave control of the government to the Republicans under Lincoln, who had made protection an important plank in the party's platform. To some extent the United States was fortunate in that many of the negative potential effects of the tariffs were somewhat offset by the free movement of capital, the large size of the internal US market, and the benefits of an extremely open immigration policy. Thus, while goods trade was restricted, capital and labour remained mostly mobile.

By the end of the 19th century, however, the free trade regime brought on by the 1860 Anglo-French Treaty began to unravel. As early as 1878 Germany began to modify her agricultural tariffs in response to pressure from farmers due to increased competition from Russia and the United States. French textile manufacturers pushed the government to abandon the treaty in 1882 and a new set of tariffs were put into place at the beginning of 1892. However, it is worth noting that in both cases the resulting tariff regimes were still relatively moderate and not comparable to the high protection of early Britain or mid-19th-century USA, and Europe still enjoyed vigorous exchange up to 1914, when the European system of open trade was effectively destroyed, first by the war and then by the high tariff walls that nations began to enact during the Great Depression.

The 19th-century trade debates have remained an important touchstone for both scholars and political elites. The same general issues persist to this day. How vigorously should a nation pursue free trade? Is it best to liberalize unilaterally or bilaterally with treaties or collectively through groups like the World Trade Organization? Today we continue to hear concerns about the importance of the trade deficit in hampering or restraining economic growth. Large and small nations often invoke the need to protect infant industries as a justification for tariffs, although it is interesting that in most cases throughout the world it is ageing and decaying industries that are likely to receive protection rather than the newer, more innovative sectors of the economy. And, as with Great Britain in the 19th century, the USA today is seen as the leader in world trade, with some of the same questions being asked about the extent to which trade is manipulated to improve world welfare or merely to enhance the narrow interests of the leading nations. And with the rise of treaties such as the North American Free Trade Agreement and the Central American Free Trade Agreement, as well as the Eurozone, there remain questions as to the virtues of piecemeal reform or the extent to which these agreements are merely mechanisms for obstructing trade by parcelling out the world into separate trading blocs.

See Also

- ▶ Globalization
- ▶ Historical Economics, British
- ▶ Smith, Adam (1723–1790)

Bibliography

Hilton, B. 1977. *Corn, cash and commerce: The economic policies of the Tory government 1815–1830*. Oxford: Oxford University Press.

Irwin, D.A. 1996. *Against the tide: An intellectual history of free trade*. Princeton: Princeton University Press.

Mun, T. 1664. *England's treasure by forraign trade*. London: Printed by J.G. for Thomas Clark.

Nye, J.V.C. 1991. The myth of free trade in Britain and fortress France: Tariffs and trade in the nineteenth century. *Journal of Economic History* 51: 23–46.

Schonhardt-Bailey, C. (ed.). 1997. *The rise of free trade. Vol. 1, Protectionism and its critics, 1815–1837*. London: Routledge.

Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations*. Oxford: Clarendon Press, 1976.

Corn Model

G. de Vivo

This expression is commonly used to denote Sraffa's interpretation of the theory of profits formulated by Ricardo in his 1815 *Essay on Profits*.

The characteristic feature of this theory is that in the production of corn there is a physical homogeneity between capital and product, because capital (which Ricardo tends to identify with the wages paid in the year) is conceived as entirely consisting of corn. Consequently the rate of profits in agriculture (production of corn) only depends upon the conditions of production of corn, and the amount of corn constituting the wage rate, and is determined independently of prices. The rate of profits of the other sectors will have to adjust to that of agriculture, by means of variations of the price of their product relative to corn. If r is the rate of profits established in agriculture, w the (corn) rate of wages, and L_i the number of workers employed in the production of commodity i, the price of i in terms of corn will be:

$$p_i = \mathrm{WL}_i(1 + r)$$

where p_i is the only unknown.

In this conception, the production of corn plays virtually the same role as the Standard system in Sraffa's *Production of Commodities* (as Sraffa himself remarks: 1960, p. 93): corn, being the only basic commodity in the system, is also the Standard commodity (only, it is not a *composite* commodity).

The expression 'corn model' is slightly misleading, in that it may easily suggest that in Ricardo's *Essay* we have a one-sector model, whereas there are in it as many sectors as in his *Principles*. (The expression might perhaps more appropriately be employed for those stylized formulations of Ricardo's theory which, though assuming the existence of another sector beside agriculture, have much in common with one-sector models: e.g. Kaldor 1955–6, pp. 211–15; Pasinetti 1960, pp. 6–10). A more appropriate expression would be that employed by Sraffa: 'corn-ratio theory' of profits (Sraffa 1951–73, I, p. xxxiii).

As Sraffa has written, the argument of the physical homogeneity of capital and product in agriculture 'is never stated by Ricardo in any of his extant letters and papers' (ibid., p. xxxi). Sraffa's reconstruction of Ricardo's argument has been questioned, in particular by S. Hollander, in his attempt to revive Marshall's interpretation of Ricardo's theory as an 'incomplete' version of marginalist theory (see Hollander 1973; 1979). It should be said, however, that Hollander does not deny the existence in Ricardo of a 'corn model'; he only denies its relevance in Ricardo's theoretical construction (Hollander 1979, p. 146; for a discussion of Hollander's arguments, see Eatwell 1975; Garegnani 1982).

A 'corn-ratio' theory of profits can be found in quite a few of Ricardo's comtemporaries (Hollander himself notices its existence in Malthus's *Measure of Value*: Hollander 1979, p. 722). The most interesting case is that of Torrens, in whose 1820 *External Corn Trade* a 'cornratio' theory is formulated much more clearly than in Ricardo. The fact that Torrens explicitly avows the Ricardian parentage of this conception, appears as a strong confirmation of Sraffa's reconstruction of Ricardo's argument (see Langer 1982; de Vivo 1985).

See Also

- British Classical Economics
- Ricardo, David (1772–1823)
- Sraffa, Piero (1898–1983)

Bibliography

De Vivo, G. 1985. Robert Torrens and Ricardo's 'corn-ratio' theory of profits. *Cambridge Journal of Economics* 9(1): 89–92.

Eatwell, J. 1975. The interpretation of Ricardo's essay on profits. *Economica* 42(166): 182–187.

Garegnani, P. 1982. On Hollander's interpretation of Ricardo's early theory of profits. *Cambridge Journal of Economics* 6(1): 65–77.

Hollander, S. 1973. Ricardo's analysis of the profit rate, 1813–15. *Economica* 40(159): 260–282.

Hollander, S. 1979. *The economics of David Ricardo*. London: Heinemann.

Kaldor, N. 1955–6. Alternative theories of distribution. *Review of Economic Studies*. As reprinted in N. Kaldor, *Essays on value and distribution*. 2nd edn, London: Duckworth. 1980.

Langer, G.F. 1982. Further evidence for Sraffa's interpretation of Ricardo. *Cambridge Journal of Economics* 6(4): 397–400.

Pasinetti, L.L. 1960. A mathematical formulation of the Ricardian system. *Review of Economic Studies*. As reprinted in L.L. Pasinetti, *Growth and income distribution: Essays in economic theory*. Cambridge: Cambridge University Press. 1974.
Sraffa, P. (ed.) 1951–73. *The Works and Correspondence of David Ricardo*, Vols I–XI. Cambridge: Cambridge University Press.
Sraffa, P. 1960. *Production of commodities by means of commodities. Prelude to a critique of economic theory*. Cambridge: Cambridge University Press.

Corporate Economy

R. L. Marris

'The Corporate Economy' is a term of art used loosely to refer to the way the economic system of rich industrialized societies evolved after 1900: a system in which a major part of the production side of the economy is organized by large limited-liability corporations whose shares are traded on organized stock markets. The phrase itself was first used openly in this sense in the title of a book edited by R. Marris and A. Wood published in 1971 (Marris and Wood 1971). Associated with it are other terms such as 'The Organizational Revolution' (Boulding 1953), 'The Managerial Revolution' (James Burnham 1941) and more recently 'The Managerial Paradigm'. The last expression is intended to convey a body of ideas amounting to a picture of the world that is both internally consistent and an agenda of questions and answers. As such, the 'managerial' paradigm is clearly distinguished both from the neo-classical paradigm and from a more loosely conceived 'neo' Marxian paradigm. It particularly depends on the proposition that in the large corporation, owing to inherent difficulties for shareholders in monitoring the performance of managers, there is considerable separation between the vicarious role of stockholders, on the one hand, and the control, operating and policy-making roles of management, on the other. This state of affairs was first identified in a classic work published in 1932 by A.A. Berle and Gardner Means (1932).

In reality, the entities involved are diverse in character, and by no means confined to 'independent' or 'private' corporations or companies. They include nationalized industries producing for the market (such representing major elements, for example, in the second half of the twentieth century in the mixed economies of Western Europe), regulated public utilities (e.g. gas or electricity supply) or even, as in Yugoslavia, fairly large 'labour-managed' entities.

Although limited-liability laws were widely promulgated in most of the countries which we now called 'western' from around 1840, by 1910 the largest hundred industrial corporations typically controlled no more than 15% of total industrial value-added (see Prais 1976). Fifty years later, the figure had trebled. In 1890, Alfred Marshall had written 'there are few exceptions to the rule that large firms . . . are, in proportion to their size, inferior to businesses of moderate size, in energy, resources . . . and inventive power' (Marshall 1980). Marshall automatically assumed that all business was family business, ignoring the potentiality of continuing organization. (See e.g. the famous passage comparing firms to 'trees of the forest' in his *Principles*, 1st edition.) Nevertheless, Marshall's characterization of industrial organization remained at the heart of neoclassical economics for the next hundred years. It assumes, almost axiomatically, that the basic economic agents on the production side of the economy, namely firms, must have decreasing returns to scale. By contrast, modern corporations attain gigantic scales with little evidence of increasing internal inefficiency. If returns to scale should be typically *increasing* or even merely constant, however, competitive equilibrium does not exist.

Institutionalist writers such as Thorsten Veblen or John Kenneth Galbraith have ascribed this internal contradiction of the neoclassical system to conspiracy, to defects of the method of economic science or to both. There is some force in either accusation.

It is not the case, however, that the problem has been studied by heterodox economists only.

Since the late 1950s, an increasing number of mainstream writers have addressed the problem. The works of Masahiko Aoki, Robin Marris, Dennis Mueller, Hiroyuki Odagiri, Edith Penrose, Herbert Simon, Hirofumi Uzawa, Oliver Williamson and George Yarrow (alphabetical, not historical, order) may all be cited. Starting from the seminal book of Edith Penrose published in 1959, continuing through the later work of Williamson, Marris and Aoki, a sub-group of these writers have emphasized the significance of *organization* in the emergence of the corporate economy.

In the economic progress of the twentieth century, in contrast to that of the nineteenth century, organization, and especially large-scale organization, has played a major role. The key factor has been the human race's ability to devise means to restrain administrative inefficiency. The proposition that provided it is properly led and organized, a large army will usually beat a smaller army, is as true in the economic as in the military sphere. The typical solution – the organizational hierarchy – is also similar in both spheres. Oliver Williamson (1970) was the first economist to emphasize the significance of the bureaucratic hierarchy in business organization, but the seed of the idea is found in an earlier contribution of Williamson's teacher, Herbert Simon (1957). (The classic conceptualizations of bureaucracy and hierarchy in modern society are, of course, due to the sociologist Max Weber.) More generally, Williamson, in a massive life-work culminating in the publication, in 1985, of his *Economic Institutions of Capitalism*, has contributed to a reconstruction of the theory of the firm to accommodate modern realities. His approach is based on transactions-analysis: the business organization, large or small, is conceived as a hidden structure of implicit contracts, existing on account of the impracticality of competitive economic organization based mainly on explicit contracts. In Williamson's conception the problem of separation of ownership from control is thus an aspect of the general 'agency problem' (control of delegatee by delegator) which then inevitably arises.

Implications of the Paradigm

From the nature of the case, the full implications of a paradigm cannot easily be summarized; if they are not diverse or unanticipated, the paradigm is not a paradigm. Questions especially addressed by the managerial paradigm are the theory of the growth of the firm; the cooperative-game theory of the firm; role of producer organizations in the stimulation of innovation and technical progress; mergers and business concentration; the micro foundations of macro economics; the theory of economic growth.

The Theory of the Growth of the Firm

In the competitive paradigm, firms are neurons of the 'invisible hand'. Like actual neurons, they are not supposed to grow, nor indeed to have any motivation of their own. But in the corporate economy, if producer organizations can and do grow indefinitely, why has not the most successful among them swallowed the whole economy? More generally, *why* and *how* do firms grow? What forces stimulate, or, alternatively, constrain, the process? Is it possible that the standard model of neoclassical micro-theory, where the absolute sizes of firms are endogenous elements in the general solution for prices and quantities in industries and in the economy, can be replaced by a theory in which a similar role is played by the rate of change of size, i.e. by the firm's rate of growth?

Penrose postulated that both stimuli and constraints for growth arose out of the firm's inherent character as an administrative organization. Marris and Uzawa postulated closed models in which the stimuli arose from the internal motives of management, the constraints from external markets for goods, money and shares. The firm could grow faster by devoting more cash-flow to growth-creating activities – such as research and development – but in so doing reduced cash-flow available for dividends. If, thus created, growth was excessive, actual or potential owners of the shares would tend to become sellers, rather than buyers, tend to depress the market price of the

stock, these results tending in turn to invite unwelcome attentions from take-over raiders. This type of model can be, and (especially in the hands of Odagiri 1981) has been, developed to determine all other variables in the theory of the firm: product prices, production technology, research effort, diversification rate and profits.

From Marris onwards, all contributors have awarded the threat of merger a major role, as indicated, in *constraining* management. Mueller (1969), however, suggested that the other side of this coin could be equally significant. Who were the typical raiders? Mueller's answer was that they would be other managerially motivated corporations seeking mergers as a means to satisfy their own ambitions. Mueller (1972) also convincingly demonstrated a 'life-cycle' version of the managerial theory, in which the greatest potentiality for divergence of interest between management and stockholders came at a late stage of the life cycle, when special opportunities were exhausted and stockholders would do better with their money released to invest elsewhere.

Cooperative Theory of the Firm

The 'managerialist' theories, such as those of Marris, Uzawa, Yarrow and Odagiri (historical order), assume that management-controlled firms are governed by managers for the benefit of managers, subject only to *constraints* from shareholders, workers or society. In sharp contrast, Aoki (1984) conceives of high management as arbitrator between the interests of existing employees and existing shareholders. This means that the high management will aim for a solution, which Aoki calls 'organizational equilibrium', that is equivalent to the solution that would be obtained if both sides were perfectly represented by agents who would behave according to the axioms of rational bargaining in game theory (see literature cited by Aoki 1984, pp. 69 et seq.). In general such processes mostly conclude by equalizing each side's proportionate utility gain as compared with the respective expected outcomes if negotiation broke down. The result represents a distribution of *organizational rent*, the latter being the surplus derived by the organization, from society, through possessing specific synergistic human resources. The employees will usually have sufficient bargaining strength to obtain some share in this rent: so the wage will exceed the 'competitive' wage. Thus in the corporate economy, as conceived by Aoki, every business organization is engaged in a general game of monopolistic competition with every other, and each will divide the resulting proceeds in some way between rewarding existing workers, rewarding existing stockholders, and growth. At the micro level, the theory, like Odagiri's, can therefore uniquely determine price, output, wages, dividends, and growth.

Formal 'New' Theories of the Firm

In the managerialist models, the firm is assumed to earn profit in a traditional way from existing activities, but to devote a part of the result to various activities intended to expand its own environment and permit long-term growth without falling profitability. The faster a firm attempts thus to grow, however, the more it is liable to become subject to what has become known as the 'Penrose effect', that is, increased costs and inefficiency caused by internal organizational problems associated with the speed of expansion. The latter costs, together with other developmental costs, are now called generically 'the costs of growth'. Normalized by reference to the size of the firm, they can be represented as an increasing function of the growth rate itself (see Eq. 1 below). For simplicity the resulting models were then formulated as steady-state growth models.

Unlike tangible investment expenditure, in business statistics costs of growth are usually treated as current expenses, deducted from cash flow before reporting profits. Reported profits are therefore operating profits less costs of growth; dividends are reported profits less cash retained for tangible investment in plant and inventory also associated with expansion. Assuming for simplicity that expansion is 100% internally financed

(it can be shown that the assumption is not important), the result (Eq. 2) is a unique relationship between the level of the dividend and the growth rate. However, in steady state, with no new external finance, if the costs of growth are satisfied, growth occurs at a constant operating profit rate, hence, with a constant number of shareholders, the growth rates of assets and of dividends per share are equal. Thus the model generates a relationship, susceptible to *managerial* choice, between level of dividends and growth of dividends. It is then assumed that current dividends and expected growth of dividends (the latter being the equivalent of anticipated capital gains), together determine stock market valuation: Eq. 3 is the standard formula for the present value of a stream of dividends growing to infinity at the rate g, discounted at the rate i, subject to the restriction $g < i$. Hence the end result is a managerial choice between growth rate and stock market value (Eq. 4), now generally known as 'the v-g frontier'.

Management is then envisaged as possessing a utility function (Eq. 5) whose arguments are, in fact, the dimensions of the frontier, the one (growth rate) representing the aspirations for future salary and psychic satisfaction, the other (valuation ratio) the aversion to risk of take-over. If management was motivated to maximize the interests of the stockholders only, caring nothing for growth per se, they would simply use Eq. 4 to maximize valuation; the resulting growth rate would be smaller but not necessarily zero.

The algebra is as follows:

$$p = a - p(g) \tag{1}$$

$$d = \{1 - r\}p = p - g \tag{2}$$

$$\upsilon = d/\{i - g\} - \{p - g\}/\{i - g\} \tag{3}$$

$$\upsilon = \{a - p(g) - g\}/\{i - g\} \tag{4}$$

$$U = U(g, v) \tag{5}$$

where p = profits per unit of assets; g = steady-state proportional growth rate of assets = proxy for growth rate of organization; a = 'operating profit rate' (p under zero growth); $p(g)$ = costs of growth ($p' > 0$); d = dividend per unit of assets; r = profit retention ratio (model assumes 100% internal financing hence $r = g/p$); υ = stock market value of equity shares, per unit of assets ('valuation ratio'); i = rate at which stock market discounts expected future earnings; U = managerial utility (g', $\upsilon' \geq 0$).

The above model was tested, with somewhat negative results, by Cubbin (1986). It has also been criticized on two major theoretical grounds, namely firstly (see Solow 1971) that it can only determine the rate of change, and not the absolute level, of the size of the organization, a weakness which is most marked in applications of micro-economic general equilibrium; secondly that the utility function, and the v-g frontier are not independent (Williamson 1966; Yarrow 1976; Odagiri 1981). The first weakness is self-evident from the equations. The second arises from the fact that the strength of the fear of take-over (which must affect the 'shape' of the managerial U-function) partly depends on the attractiveness of the firm to outsiders, and the latter in turn depends on the maximum value that could be extracted under valuation-maximizing policies – i.e. on the location of the peak of the v-g frontier. In their different contexts, both flaws are serious.

Odagiri (1981) escapes both weaknesses first by marrying the model to a conventional static theory of the firm, with fully specified conventional elements (production function, factor prices etc.) and then by reconstructing the managerial utility-maximizing procedure. For the second stage, he assumes that (a) there is an exogenous 'cost of take over', i.e. premium over current market share price which, owing to capital-market imperfection, must be paid by any raider and (b) that management maximizes discounted expected future salary subject to the hazard of being taken over, i.e. subject to the hazard of employment and salary terminating entirely. As described below, Odagiri's further, major, achievement consisted of aggregating from the micro level, thus modified, to a macro-growth theory.

Similar algebra can also be used to illustrate Aoki's theory. In reviewing Aoki (1984) in a Japanese journal in 1986/1987, Marris suggested

that the real wage in the Aoki theory would become,

$$w = 1/z + s(z - 1) \tag{6}$$

where $z = 1–1/e$, and $e =$ monopolistic competition elasticity of demand. From which it follows (assuming constant returns and no non-labour non-capital inputs) that the operating profit rate is given by (7),

$$a = \{k - w\}/c \tag{7}$$

where $k =$ output per worker and $c =$ capital output ratio. Hence

$$a = f(k, c, z, s). \tag{8}$$

If (8) is substituted into (4) and U re-maximized subject to the reconstituted constraint it seems intuitively likely that the optimum growth rate, and the workers' bargaining share (s) will be negatively related. Aoki confirms rigorously that this is so (1984, pp. 78 et seq). Thus intuitive 'marriage' of Odagiri (see below) and Aoki suggests that other things being equal a corporate economy will grow faster (i) the less the workers' bargaining power; (ii) the less the shareholders' bargaining power (as reflected in take-over conditions); (iii) and the less the competitiveness of the economy as reflected in elasticity of demand! These are classic examples of conclusions from the managerial paradigm that cannot be obtained from the neoclassical paradigm.

General Micro-economics

At the end of the day, the concept of the 'industry' becomes blurred, and the production side of the economic stage becomes inhabited by agile dinosaurs, 'competing' for growth and profits but not necessarily engaging in strong price competition in individual markets. If the domestic economy becomes too small for them, they become multinational. Following a scheme suggested by Marris (1971), we may see an individual economy in this universe as a two-way table in which rows represent named organizations, columns, products or 'markets'. (In the Marshallian terminology, the rows would be 'firms', the columns 'industries'.) Elements in this matrix represent the contribution of a given organization to a given market. In the neoclassical paradigm, there is only one entry in each row (one firm, one industry) and many entries per column (perfect competition). In the corporate economy, there may be a number of entries in a row (any organization may be diversified); but, on account of economies of scale and the instability of competitive equilibrium, there will usually be only a moderate number of entries in columns, i.e. the typical situation is oligopoly.

It follows that the micro-economics of the corporate economy are essentially based on theories of imperfect competition and oligopoly; on this analysis the competitive economy is impossible.

But the matrix is not static. A major part of managerial effort is devoted to efforts to changing it. As a result of R&D, the list of columns (products) changes. As a result of mergers, the list of rows (organizations) changes. The general process needs analyzing by the micro theory of the growth of the firm, and, in turn, contributes to the micro foundations of macro growth economics.

Economic Concentration

In industrial economics it is usual to assess the potential competitiveness of a market by two factors, the proportion of sales controlled by the largest n firms (where n is a number like 4 or 5), and other elements in barriers to entry. In our matrix, 'industrial concentration' is therefore a characteristic of the columns. Entry occurs in the rows; its most likely cause is a previously zero element becoming positive; an existing organization enters a new field. Since existing organizations may well be large, the corporate economy does not reduce, and may increase, this type of competition. Nor does this system necessarily imply any long term tendency for increased average 'column-wise' (i.e. 'within-industry') concentration.

But another form of economic concentration occurs in the *row-totals*. Sometimes known as

'business concentration' or 'macro concentration', this type can be measured by the proportion of industrial output controlled by the *m* largest organizations, where *m* is a number like 100 or 200. That statistic increased sharply in all countries between 1910 and 1960, and is still apparently increasing in Europe. In the US, in the past 20 years, it has apparently been stabilized by new entry into the total organizational population (see Spillberg 1985).

Macro concentration will be the product of two forces: (i) the internal growth of firms (organizations); (ii) mergers. Building on the earlier work of Gibrat (1931) and Prais (1976) developed an elegant application of the theory of Markov chains to show how stochastic disturbance of internal growth rates could produce an automatic tendency towards concentration without invoking any other systematic forces. Marris (1979) married this theory with the theory of the growth of the firm, and also with a managerialist/stochastic model of the merger process, to produce a general theory of concentration. If there are constant returns to scale, in the absence of macro new entry, business concentration will increase at a constant rate through time. With increasing returns, the process is exploding. With decreasing returns, concentration increases at a diminishing rate converging to an asymptote. (Although decreasing returns in respect of absolute size is supposedly ruled out in the managerial paradigm, the same results are also obtained if there are corresponding associations between absolute size and rate of change of size.)

Micro Foundations of Macro Growth Economics

One of the weaknesses of macro-economic growth theory, as it emerged in the aftermath of the Keynesian revolution, lay in depending on a general rate of technological progress which was not only exogenous but unexplained. An advantage of corporate-economy theory lies in being able to explain the growth rate, at least in part, from *within the model*. The proposition was hinted at in the last chapter of Marris (1964) but was not further developed before the appearance of an extremely comprehensive theory published by H. Odagiri in 1981. After refining and synthesizing the micro theory of the growth of the firm, Odagiri elegantly succeeds in aggregating to the economy-wide level, producing a model in which a number of interpretable behavioural or technical variables, together with government-policy variables, such as monetary and fiscal variables, actually determine the growth rate of the economy. The crucial factor is that, in striving for growth of *itself*, an organization *creates* (via induced technical progress) growth for the economy. Two micro factors are keys to Odagiri's macro growth rate; the strength of management's risk aversion when trading fear of take-over with desire for growth-led salary increase; and the institutional ease or difficulty of take-overs. In the second half of the twentieth century, in Japan, on the one hand, under the system of 'life time employment' (informal security of tenure) managers, effectively locked into their firms, were in consequence likely to have comparatively strong growth preference, while on the other hand the organization of the Japanese capital market was not conducive to ease of take-overs. From such facts the managerial theory would predict fast growth of firms and the economy. The neoclassical theory, by contrast, would either have nothing to say, or give the opposite predictions. The actual result was that during this period, the growth rate of Japanese economy was by any standard exceptional; the 'Japanese miracle' may thus be seen as an outstanding achievement of the Corporate Economy.

See Also

- ▶ Economic Theory of the State
- ▶ Industrial Organization

References

Alchian, A.A. and H. Demsetz. 1972. Production, information costs and economic organization. *American Economic Review* 62: 777–795.

Aoki, M. 1984. *The co-operative game theory of the firm*. Oxford: Oxford University Press.
Boulding, K.J. 1953. *The organizational revolution*. New York: Harper.
Burnham, J. 1941. *The managerial revolution*. New York: The John Day Co.
Cubbin, J. 1986. Testing the Marris model. *Managerial Economics*.
Galbraith, J.K. 1952. *American capitalism*. Boston: Houghton Mifflin.
Galbraith, J.K. 1967. *The new industrial state*. Boston: Houghton Mifflin.
Gibrat, F. 1931. *Les inégalités économiques*. Paris: Recueil Sirey.
Kuehn, D. 1975. *Takeovers and the theory of the firm*. London: Macmillan.
Marris, R.L. 1964. *The economic theory of 'Managerial' capitalism*. London: Macmillan.
Marris, R.L. 1979. *The theory and future of the corporate economy*. Amsterdam: North-Holland Press.
Marris, R.L. 1986/1987. Review of Aoki (1984) in *Keizai Kenko*.
Marris, R.L. and Adrian Wood. 1971. *The corporate economy*. London: Harvard University Press.
Marshall, A. 1890a. *Address* to the British Association.
Marshall, A. 1890b. *Principles of economics*. Successive editions, 1891–1920. London: Macmillan.
Means, G.C. and A.A. Berle. 1932. *The modern corporation and private property*. New York: Macmillan.
Mueller, D. 1969. A theory of conglomerate merger. *Quarterly Journal of Economics* 83(4): 643–659.
Mueller, D. 1972. A life cycle theory of the firm. *Journal of Industrial Economics* 20(3): 199–219.
Mueller, D. 1984. Further reflections on the invisible-hand theorem. In *Economics in disarray*, ed. P. Wiles. Oxford: Oxford University Press.
Mueller, D., and R.L. Marris. 1980. The corporation, competition and the invisible hand. *Journal of Economic Literature* 18: 32–63.
Odagiri, H. 1981. *The theory of growth in a corporate economy*. Cambridge: Cambridge University Press.
Odagiri, H. 1982. Anti-neoclassical management motivation in a neoclassical economy: A model of economic growth and Japan's experience. *Kyklos* 35(2): 223–243.
Penrose, E. 1959. *The theory of the growth of the firm*. Oxford: Oxford University Press.
Penrose, E. 1985. *The theory of the growth of the firm 25 years after*. Uppsala: Acta Universitatis Upsalienis.
Prais, S. 1976. *The evolution of giant firms in Britain*. Cambridge: Cambridge University Press.
Simon, H. 1957a. *Models of man*. New York: Wiley.
Simon, H. 1957b. The compensation of executives. *Sociometry* 20(1): 32–35.
Simon, H. 1972. Theories of bounded rationality. In *Decision and organization*, ed. C. Maguire and R. Radner. Amsterdam: North-Holland.
Simon, H. and C. Bonini. 1958. The size distribution of business firms. *American Economic Review* 48: 607–617.
Solow, R. 1971. Some implications of alternative criteria for the firm. In Marris and Wood (1971).
Uzawa, H. 1969. Time preference and the Penrose effect in economic growth. *Journal of Political Economy* 77: 628–652.
Veblen, T. 1923. *Absentee ownership and business enterprise in recent times*. New York: B.W. Huebsch.
Weber, M. 1921. In *Economy and society*, ed. G. Roth. New York: Bedminster Press, 1968.
Williamson, O. 1964. *The economics of discretionary behavior*. Englewood Cliffs: Prentice Hall.
Williamson, J. 1966. Profit, growth and sales maximization. *Economica* 33: 1–16.
Williamson, O. 1970. *Corporate control and business behavior*. Englewood Cliffs: Prentice Hall.
Williamson, O. 1975. *Markets and hierarchies*. Englewood Cliffs: Prentice Hall.
Williamson, O. 1985. *The economic institutions of capitalism*. New York: Free Press.
Yarrow, G. 1976. On the predictions of managerial theories of the firm. *Journal of Industrial Economics* 24(4): 267–279.

Corporate Governance

Luigi Zingales

Abstract

Introduced in the mid-1980s, the term 'corporate governance' can be defined as the set of conditions that shapes the *ex post* bargaining over the quasi-rents generated by a firm. The incomplete contracts approach has been very successful in explaining the corporate governance of entrepreneurial firms and also some important features of large corporations, such as allocation of ownership to the providers of capital who are dispersed, and the importance of internal organization. Aspects that remain to be investigated include the role of the board of directors, interaction between different mechanisms of corporate governance, and the normative implications of the approach.

Keywords

Asymmetric information; Bargaining; Bilateral monopoly; Coase theorem; Contractual

governance; Control rights; Corporate governance; Firm, theory of; Free-rider problem; Human capital; Incomplete contracts; Mechanism design; Nexus of contracts view of the firm; Property rights view of the firm; Quasi-rent; Risk aversion; Takeovers

JEL Classifications
G30; P50

While some of the questions have been around since Berle and Means (1932), the term 'corporate governance' did not exist in the English language until the mid-1980s. Since then, however, corporate governance issues have become important not only in the academic literature but also in public polity debates. During this period, corporate governance has been identified with takeovers, financial restructuring and institutional investors' activism. But what exactly is corporate governance? Why is there a corporate governance 'problem'? Why does Adam Smith's invisible hand not automatically provide a solution? What role do takeovers, financial restructuring and institutional investors play in a corporate governance system?

In this article I will try to provide a systematic answer to these questions, making explicit the essential link between corporate governance and the theory of the firm. My goal is to provide a common framework that helps to analyse the results obtained in these two fields and identify the questions left unanswered. This is not a survey, so I make no attempt to be comprehensive. For an excellent survey on the topic the reader is referred to Shleifer and Vishny (1997).

When Do We Need a Governance System?

The word 'governance' is synonymous with the exercise of authority, direction, and control. These words, however, seem strange when used in the context of a free- market economy. Why do we need any form of authority? Isn't the market responsible for allocating all resources efficiently without the intervention of authority? The basic (neoclassical) undergraduate microeconomics courses rarely mention the words 'authority' and 'control'.

In fact, neoclassical microeconomics describes well only one set of transactions, which Williamson (1985) calls 'standardized'. Consider, for instance, the purchase of a commodity, like wheat. There are many producers of the same quality of wheat and many potential customers. In this context, Adam Smith's invisible hand ensures that the good is provided efficiently without the need of any form of authority.

Many daily transactions, however, do not fit this simple example. Consider, for instance, the purchase of a customized machine. The buyer must contact a manufacturer and agree upon the specifications and the final price. Unlike the case of wheat, the signing of the agreement does not represent the end of the relationship between the buyer and the seller. Producing the machine requires some time. During this time many events can occur, which alter the cost of producing the machine as well as the buyer's willingness to pay for it. More importantly, before the agreement was signed, the market for manufacturers was competitive. Once production has begun, though, the buyer and the seller are trapped in a situation of bilateral monopoly. The customized machine probably has a higher value to the buyer than to the market. On the other hand, the contracted manufacturer has probably the lowest cost, to finish the machine. The difference between what the two parties generate together and what they can obtain in the marketplace represents a quasi-rent, which needs to be divided *ex post.* In dividing this surplus Adam Smith's invisible hand is of no help, while authority does play a role.

In the spirit of Williamson (1985), I define a *governance system* as the complex set of constraints that shape the *ex post* bargaining over the quasi-rents generated in the course of a relationship. A main role in this system is certainly played by the initial contract. But the contract, most likely, will be incomplete, in the sense that it will not fully specify the division of surplus in every possible contingency (this might be too costly to do or outright impossible because the

contingency was unanticipated). This creates an interesting distinction between decisions made *ex ante* (when the two parties entered a relationship and irreversible investments were sunk) and *ex post* (when the quasi-rents are divided). This contract incompleteness also creates room for bargaining.

The outcome of the bargaining will be affected by several factors besides the initial contract. First, which party has ownership of the machine while it is being produced. Second, the availability of alternatives: how costly is it for the buyer to delay receiving the new machine; how costly is it for the manufacturer to delay the receipt of the final payment; how much more costly is it to have the job finished by another manufacturer, and so on. Finally, a major role in shaping the bargaining outcome is played by the institutional environment. For example: how effective and rapid is law enforcement; what are the professional norms; how quickly and reliably does information about the manufacturer's performance travel across potential clients, and so on. All these conditions constitute a governance system.

As illustrated by the machine example, there are two necessary conditions for a governance system to be needed. First, the relationship must generate some quasi-rents. In the absence of quasi-rents, the competitive nature of the market will eliminate any scope for bargaining. Second, the quasi-rents are not perfectly allocated *ex ante.* If they were, then there would be no scope for bargaining either.

Corporate Governance

The above definition of governance is quite general. One can talk about the governance of a transaction, of a club, and, in general, of any economic organization. In a narrow sense, corporate governance is simply the governance of a particular organizational form – a corporation.

Yet the bargaining over the *ex post* rents, which I defined as the essence of governance, is influenced, but not uniquely affected, by the legal structure used. A corporation, in principle, is just an empty legal shell. What makes a corporation valuable is the claims the legal shell has on an underlying economic entity, which I shall refer to as the firm. While often the legal and the economic entity coincide, this is not always the case. For this reason, I define *corporate governance* as the complex set of constraints that shape the *ex post* bargaining over the quasi-rents generated by a firm.

To be sure, many problems that fall within the realm of corporate governance can be (and have been) profitably analysed without necessarily appealing to such a broad definition. Nevertheless, all the governance mechanisms discussed in the literature can be reinterpreted in light of this definition. Allocation of ownership, capital structure, managerial incentive schemes, takeovers, boards of directors, pressure from institutional investors, product market competition, labour market competition, organizational structure, and so on can all be thought of as institutions that affect the process through which quasi-rents are distributed. The contribution of this definition in simply to highlight the link between the way quasi-rents are distributed and the way they are generated. Only by focusing on this link can one answer fundamental questions such as who should control the firm.

Of course, this definition of corporate governance raises the age-old question of what a firm is. But this question should be central to corporate governance. Before we can discuss how a firm should be governed, we need to define the firm. This question is also important because it helps us identify to what extent, if any, corporate governance is different from the governance of a simple contractual relationship (such as in the machine example).

There are two main definitions of the firm available in the literature. The first, introduced by Alchian and Demsetz (1972), is that the firm is a nexus of contracts. According to this definition, there is nothing unique in corporate governance, which is simply a more complex version of standard contractual governance.

The second definition, due to Grossman and Hart (1986) and Hart and Moore (1990) (henceforth GMH), is that the firm is a collection of physical assets that are jointly owned. Ownership matters because it confers the right to make

decisions in all the contingencies unspecified by the initial contract. On the one hand, this definition has the merit of differentiating between a simple contractual relationship and a firm. Since the firm is defined by the non-contractual element (that is, the allocation of ownership), corporate governance (as opposed to contractual governance) is defined by the effect of this non-contractual element. Not surprisingly, the focus of the corporate governance literature since the mid-1990s has been the allocation of ownership (hence this literature is called the property rights view of the firm). On the other hand, this definition has the drawback of excluding any stakeholder other than the owner of physical assets from being important to our understanding of the firm.

More recently, Rajan and Zingales (2001, 1998) have proposed a broader definition. They define the firm as a nexus of specific investments: a combination of mutually specialized assets and people. Unlike the nexus of contracts approach, this definition explicity recognizes that a firm is a complex structure that cannot be instantaneously replicated. Unlike the property rights view, this definition recognizes that all the parties who are mutually specialized belong to the firm, be they workers, suppliers or customers. While this definition does not necessarily coincide with the legal definition, it does coincide with the economic essence of a firm: a network of specific investments that cannot be replicated by the market.

Incomplete Contracts and Governance

In an Arrow–Debreu economy it is assumed that agents can costlessly write all state- contingent contracts. As a result, all decisions are made *ex ante* and all quasi-rents are allocated *ex ante*. Thus, there is no room for governance. More surprisingly, even if we relax the assumption that every state-contingent contract can be written and admit that certain future contingencies are not observable (and thus not contractible), we still find no room for governance as long as one can costlessly write contracts on all future observable variables.

Recall the example of the customized machine, and assume that the manufacturer's effort is unobservable to others and is, therefore, not contractible. The neoclassical approach to this problem is to design a mechanism (hence the term 'mechanism design'), contingent on all publicly observable variables, which provides the manufacturer with the best possible incentives to exert effort. Myerson (1979) shows that all optimal mechanisms are equivalent to a revelation (direct) mechanism in which the agent (manufacturer) publicly announces his information and receives compensation contingent on his announcement. An important consequence of this result is that, in the mechanism design approach, delegation (giving an agent discretion over certain decisions) is always weakly dominated by a fully centralized mechanism, where all decisions are made *ex ante* by the designer. The mechanism design approach reproduces several distinguishing features of an Arrow–Debreu economy: all decisions are made *ex ante* and executed only *ex post*; as a result, all conflicts are resolved and all rents are allocated *ex ante.* This leaves no room for *ex post* bargaining. All these features are incompatible not only with my definition of governance, but also with any meaningful (that is, related to the sense in which this term has been used) definition of governance. This is best illustrated with two examples.

One of the crucial questions in corporate governance concerns the party in whose interest corporate directors should act. In the mechanism design approach this question cannot even be raised. All possible future conflicts are resolved *ex ante* and the initial contract specifics how directors will behave in any observable state of the world. However, since this question is raised all the time, it must be that all possible conflicts are not resolved *ex ante.*

Second, the mechanism design approach avoids renegotiation: the initial contract is so designed that the agents do not want to renegotiate. As a result, the designer wants to make renegotiation as inefficient as possible: this reduces the costs of providing incentives to the agents with no efficiency costs, since renegotiation never occurs in equilibrium (Aghion et al. 1997).

If this result were to be taken seriously, the optimal public policy approach would be to preserve any inefficiency in the system in order to avoid destroying its beneficial incentives *ex ante.* In reality, though, the jurisprudential approach is completely different. For example, courts do not support punitive damages that are considered excessive with respect to the issue at stake.

Only in a world where some contracts contingent on future observable variables are costly (or impossible) to write *ex ante* is there room for governance *ex post.* Only in such a world are there quasi-rents that must be divided *ex post* and real decisions that must be made. Finally, only in a world of incomplete contracts can we define what a firm is and discuss corporate governance as being different from contractual governance. Not surprisingly, the theory of governance is intimately related to the emergence and evolution of the incomplete contracts paradigm.

A fundamental milestone in this evolution is the residual rights of control concept introduced by Grossman and Hart (1986). In a world of incomplete contracts, it is necessary to allocate the right to make *ex post* decisions in unspecified contingencies. This residual right is both meaningful and valuable. It is meaningful because it confers the discretion to make decisions *ex post*. It is valuable because this discretion can be used strategically in bargaining over the surplus.

Why Does Corporate Governance Matter?

By definition, corporate governance matters for distribution of rents, but to what extent does it matter for economic efficiency? There are three main channels through which the conditions that affect the division of quasi-rents also affect the total surplus produced. In presenting these channels I make a sharp distinction between *ex ante* (when specific investments need to be sunk) and *ex post* (when quasi-rents are divided) effects, as though the firm lasted just one period. Of course, this is not true in reality because *ex post* considerations of one period are mixed with *ex ante* considerations for the next period.

Ex Ante *Incentive Effects*

The process through which surplus is divided *ex post* affects the *ex ante* incentives to undertake some actions, which can create or destroy some value, in two main ways.

First, rational agents will not spend the optimal amount of resources in value- enhancing activities that are not properly rewarded by the governance system. In fact, one goal in designing a governance system is to motivate those investments that are not properly rewarded in the marketplace. The canonical example of how a change in the governance structure can change the incentives to make a value- enhancing relationship-specific investment is the Fisher Body case. In the early 1920s, Fisher Body (an auto body manufacturer) refused to locate its plants close to General Motors' plants in spite of the obvious efficiency improvement generated by such a move. Locating close to GM would have reduced Fisher Body's ability to supply other car manufacturers, which would have weakened its bargaining position *ex post* and possibly reduced its share of the quasi-rents generated by the relationship with GM (see Klein et al. 1978). A change in the governance system (the acquisition of Fisher Body by GM) eventually led to the efficient plant location decision. Another famous illustration of the same phenomenon is managerial shirking. A manager will shirk if her *ex post* bargaining payoff does not increase sufficiently with her effort and, therefore, fails to compensate her for the cost of this effort.

Second, rational agents will spend resources in inefficient activities whose only (or main) purpose is to alter the outcome of the *ex post* bargaining in their favour.

For example, a manager may specialize the firm in activities she is best at running because this increases her marginal contribution *ex post* and, thus, her share of the *ex post* rents (Shleifer and Vishny 1989). Interestingly, this problem is not limited to the top of the hierarchy, but is present throughout. Subordinates, who do not have much decision power, will waste resources trying to capture the benevolence of their powerful superiors (Milgrom 1988). Even the well-known tendency of managers to overinvest in growth can be interpreted as a manifestation of

this problem. Managers like to expand the size of their business because this makes them more important to the value of the firm and, thus, increases the payoff they can extract in the *ex pest* bargaining.

Of course, a governance system might promote or discourage these activities. For example, Chandler (1966) reports that, under the Durand reign, GM's capital allocation was highly politicized ('a sort of horse trading'). The move to a multidivisional structure, with the resulting increase in divisional managers' autonomy, reduced the managers' payoff from rent-seeking. Similarly, Milgrom and Roberts (1990) explain many organizational rules as a way to minimize influence costs. Finally, Rajan et al. (2000) argue that inefficient 'power-seeking' is more severe the more investment opportunities a firm's divisions have. Consistent with this claim, they find that the value of a diversified firm is negatively related to the diversity of the investment opportunities of its divisions.

Thus, a governance system affects the incentives to invest or seek power, altering the marginal payoffs that these actions have in *ex post* bargaining. For instance, for an independent Fisher Body, the marginal effect on the bargaining payoff of localizing its plants close to GM is negative (it reduces the value of its outside options), but is positive for Fisher Body as a unit of GM, which does not have the authority to supply other manufacturers without GM's consent (see Rajan and Zingales 1998). Thus, a different ownership structure alters the incentives to make specific investments.

Inefficient Bargaining

A second channel through which a governance system affects total value is by altering *ex post* bargaining efficiency. This is tantamount to saying that the governance system affects the degree to which the assumptions of the Coase theorem are violated. A governance system, therefore, can affect the degree of information asymmetry between the parties, the level of coordination costs, or the extent to which a party is liquidity constrained.

For example, if control rights are assigned to a large and dispersed set of claimants (like the shareholders of most publicly traded companies), free-rider problems may prevent an efficient action from being undertaken even if property rights are well defined and perfectly tradeable (Grossman and Hart 1980). Alternatively, the allocation of control rights can affect efficiency by determining the direction in which a compensating transfer must be made. The direction of the transfer matters when one of the parties to the *ex post* bargaining is liquidity constrained (Aghion and Bolton 1992) or when it faces a different opportunity to invest these resources productively rather than in power-seeking activities (Rajan and Zingales 1996). In both cases an efficient transaction may not be agreed upon – in the first case because the party that should compensate does not have the resources, in the second case because the transaction (while efficient *per se*) may generate such an increase in wasteful power-seeking as to more than offset its benefits.

To this standard list of imperfections, Hansmann (1996) adds the divergence of interests among the parties who have control rights. Citing the political economy literature, Hansmann argues that *ex post* inefficiency is increasing in the divergence of interests among control holders. For example, he argues that allocating control to workers is more costly when they differ in their professional skills, hierarchical position and tenure. While Hansmann does not provide a formal model of why this relation occurs, he does provide very compelling evidence that in practice control rights are rarely allocated to parties with conflicting Interests. His conjecture is intriguing because there is no well-established general theory of how different governance systems lead to different levels of *ex post* inefficiency. There is little doubt, however, that these inefficiencies exist and are important. For example, Wiggins and Libecap (1985) document that an excessively dispersed initial allocation of drilling rights leads to an inefficient method of extracting oil, with estimated losses as big as 50 per cent of the total value of the reservoir.

Risk Aversion

Finally, a governance system might affect the *ex ante* value of the total surplus by determining the level and the distribution of risk. If the different parties have different degrees of risk aversion (or different opportunities to diversify or hedge risk), then the efficiency of a governance system is also measured by how effectively it allocates risk to the most risk-tolerant party. This idea is the cornerstone of Fama and Jensen's (1983a; 1983b) analysis of organizational structure and corporate governance.

Different governance systems can also *generate* a different amount of risk. Suppose, for instance, that the total amount of surplus generated is constant. It is still possible that the payoff of each party is stochastic, if the governance structure generates a stochastic bargaining outcome. For example, a life insurance contract written in nominal dollars creates a pure gamble between the policy holders and the insurance company with respect to the future rate of inflation. This additional 'governance' risk (in this case created by the contract, in general created by the governance structure) reduces the value of the total surplus, if the parties are risk averse and cannot diversify away the risk.

In summary, the objectives of a corporate governance system should be: (*a*) to maximize the incentives for value-enhancing investments, while minimizing inefficient power seeking; (*b*) to minimize inefficiency in *ex post* bargaining; (*c*) to minimize any 'governance' risk and allocate the residual risk to the least risk averse parties.

Who Should Control the Firm?

To show the utility of the framework developed thus far, I will use it to address one of the most controversial issues in corporate governance: who should control the firm? In particular, I will analyse it with regard to the first of the three above objectives of a corporate governance system. For an analysis focused on the second objective the reader is referred to Hansmann (1996), and for an analysis focused on the third to Fama and Jensen (1983a, b).

As far as the first objective is concerned, the allocation of control is important because it affects the division of surplus. By controlling a firm's decisions, a party can ensure for itself of more and more valuable options without the collaboration of the other parties. This guarantees the controlling party a larger share of the surplus within the relationship. Thus, in the framework outlined above, the question of who should control the firm can be rephrased as: whose investments need more protection in the *ex post* bargaining? Again, the answer to this question is indissolubly linked to the underlying theory of the firm.

In the nexus of contracts view, the firm 'is just a legal fiction which serves as a focus for the complex process in which the conflicting objectives of individuals . . . are brought in equilibrium within a framework of contractual relationship' (Jensen and Meckling 1976, p. 312). Thus, according to this view each party is fully protected by its contract with the exception of the shareholders, who accept a residual payoff because they possess a comparative advantage in diversifying risk.

As a result, shareholders need the protection insured by control.

While widely popular, this explanation is unsatisfactory. The contractual protection provided to the parties involved in the nexus of contracts is complete only if contracts are complete. But if contracts are complete, then the statement that shareholders are in control is meaningless. In fact, in a world of complete contracts all the decisions are made *ex ante*, and thus shareholders are no more in control than are the workers: everything is contained in the initial grand contract. Furthermore, as I have already argued in Section 3, this conclusion is inconsistent with the existence of a debate on what a company should do.

Alternatively, if contracts are incomplete, then the argument that all other parties are fully protected by their contractual relationships does not automatically follow. In fact, in this context one should ask why shareholders need more protection than other parties to the nexus of contracts. I return to this issue below.

In the property rights view of the firm, the reason why shareholders should be in control is straightforward. Control is allocated so as to maximize the incentives to make human capital-specific investments. The owner of the firm will generally be the worker with the most expropriable investment. In other words, the property rights approach does not deal with outside shareholders and, thus, it applies only to entrepreneurial firms.

Outside of the GHM framework, the typical justification for why shareholders (or more generally the providers of finance) are in control is based on a combination of three arguments. Shareholders need more protection because: (a) their investment is more valuable; (*b*) other stakeholders can protect their investments better through contracts; (*c*) other stakeholders have other sources of power *ex post* that protect their investments. Of the three arguments, the first is clearly unfounded. Reviewing the empirical evidence on the return on specialized human capital, Blair (1995) estimates that the quasi-rents generated by specialized human capital are as big as accounting profits, which are likely to overestimate the quasi-rent generated by physical capital. Hence, there is no ground for dismissing human capital investments as second order to physical capital investments.

The second argument is harder to dismiss. Since we lack a fully satisfactory theory of why contracts are incomplete, we cannot easily argue which contracts are more incomplete. Nevertheless, it is hard to argue that human capital investments are easier to contract than physical capital investments. If there is one contingency that is easily verifiable, it is the provision of funds. Thus, it is not obvious why providers of funds are at a comparative disadvantage.

The most convincing argument is probably the third. As Williamson (1985) puts it,

> the suppliers of finance bear a unique relation to the firm: The whole of their investment in the firm is potentially placed at hazard. By contrast, the productive assets (plant and equipment; human capital) of suppliers of raw material, labor, intermediate product, electric power, and the like normally remains in the suppliers' possession.

Thus, the other stakeholders have a better outside option in the *ex post* bargaining, and they do not need the protection ensured via the residual rights of control.

Even this argument, however, is not fully satisfactory. In fact, it suggests only that the suppliers of finance should have some form of contractual protection – it does not necessarily imply that they should be protected via the residual rights of control.

A satisfactory explanation of why the residual right belongs to the shareholders can be obtained only in a theory of the firm that explicitly accounts for the existence of different stakeholders, and models the interaction between contractual (for example, ownership) and non-contractual (for example, unique human capital investments) sources of power. An attempt in this direction is made by Rajan and Zingales (1998).

To understand the argument, note that the residual right of control over an asset always increases the share of surplus captured by its owner (who has the opportunity to walk away with the asset), but it does not necessarily increase her marginal incentive to specialize. If, as is likely, a more specialized asset has less value outside the relationship for which it has been specialized, then specialization decreases the owner's outside opportunity and,; thus, her share of the quasi-rents. Owning a physical asset, then, makes an agent more reluctant to specialize it. As a result, the residual right of control is best allocated to a group of agents who need to protect their investment against *ex post* expropriation, but who have little control over how much the asset is specialized.

Consider now the different members of the specific investments nexus that makes up the firm. Most of the specific investments which form this nexus are in human capital and, therefore, can neither be contracted nor delegated *ex ante.* Granting the residual right to any of these members will have a negative effect on their incentive to specialize. By contrast, since the provision of funds is easily contractible, funds will be provided in the optimal amount as long as their providers receive sufficient surplus *ex post*. Thus, allocating the residual rights of control to the providers of funds has the positive effect of granting them enough surplus *ex post*, while avoiding the

negative effect of reducing their marginal incentives.

Once they have provided funds, however, financial investors might be reluctant to use these funds for very specialized projects, for fear of seeing their share of the return fall. Thus, it is optimal that, while retaining a residual right of control over the assets, the providers of funds delegate the right to specialize the assets to a third party, who does not internalize the opportunity loss generated by this specialization. This third party, thus, should not be in the position of a mere agent, who owes a duty of obedience to the principal, but should be granted the independence to act in the interest of the firm (that is, the whole body of members of the nexus of specific investments), and not only of the shareholders. Blair and Stout (1997) claim that this is the role American corporate law attributes to the board of directors.

In sum, a broader definition of the firm allows us to understand why the residual right of control is allocated to the providers of capital and why its use is mostly delegated to a board of directors.

Normative Analysis

An interesting, and largely unexplored, application of the incomplete contract approach to corporate governance is the analysis of its normative implications. In a world of complete contracts, such analysis has limited scope. A benevolent social planner would be unable to improve the *ex ante* allocation reached by private contracting, because this will achieve the constrained-efficient outcome. *Ex post*, the outcome might be inefficient, but that inefficiency is always part of the written contract and needs to be preserved to maintain *ex ante* future efficiency. By contrast, in a world of incomplete contracts, there is ample scope to analyse both *ex ante* and *ex post* efficiency.

First, a privately optimal governance system may not be socially efficient. In fact, a world of incomplete contracts generates some incentives to 'arbitrage power' through time. Consider an entrepreneur, who has immense bargaining power today, but anticipates losing it in the near future. If she could write all the contracts she could succeed in extracting all the present and future surplus arising from a relationship without any distortion. But, if some contracts cannot be written, then the entrepreneur has an incentive to distort her choices so as to transfer some of her bargaining power today into the future, enabling her to capture some of the future surplus as well. This is the idea underlying the choice of ownership in Zingales (1995a) and Bebchuk and Zingales (1996), and of the hierarchical structure in Rajan and Zingales (2001). It can also be used to provide a rationale for the existing mandatory rules (see Bebchuk and Zingales 1996).

Second, in a world of incomplete contracts one can discuss the welfare effects of different institutions. For example, in a world of complete contracts the type of legal system a country adopts is irrelevant, as long as private contracts are enforced. By contrast, it is at least conceivable that in an incomplete contract world it may have a significant effect. This is intriguing because empirically it has been shown that legal institutions have an effect on the appropriability of quasi-rents by outside investors (Zingales 1995b), on the way corporate governance is structured (La Porta et al. 1996), and on the amount of external finance raised (La Porta et al. 1997).

Finally, the incomplete contract approach generates a potential role for government intervention *ex post*. Unlike in the mechanism design approach, in an incomplete contract world *ex post* inefficiency is not necessarily desirable *ex ante.*

Thus, a selective intervention that eliminates *ex post* inefficiency, while preserving the distributional consequences sought *ex ante,* will improve welfare.

Limitations of the Incomplete Contract Approach

While the incomplete contracts approach to corporate governance has brought tremendous insights to the corporate governance debate, it has two weaknesses.

First, its predictions for the optimal allocation of ownership are extremely sensitive to what

contracts can be written. Consider, for instance, the plant localization problem discussed above. If no contracts can be written, then – according; to the property rights approach – Fisher Body (who makes the bigger specific investment) should own the asset. However, if General Motors could credibly commit through a contract to buy all its car bodies from Fisher Body (as it did), then giving ownership to Fisher Body will confer too much power on it, and, thus, it is optimal for General Motors to own the asset (Hart 1989), Thus, who should have the residual right of control depends crucially upon what the contractable rights are. But this is very difficult to argue on a priori grounds without a general theory of why contracts are incomplete (see Maskin and Tirole 1997).

Second, this approach relies heavily (as does the complete contract approach) on the agents anticipating all future possible contingencies (Hellwig 1997). This requirement can be reasonable when the subject of analysis is a small entrepreneurial firm, but it loses credibility when it is applied to large publicly held companies formed decades ago. Can we really interpret the capital structure of IBM today as the outcome of the design by Charles Flint (its founder) in 1911 attempting to allocate control optimally? Hart (1995) argues that the 'founding father' interpretation is simply a metaphor for the capital structure that a manager will choose under the pressure of the corporate control market. Yet Novaes and Zingales (1995) show that the two approaches lead to different predictions, not only about the level of debt but also about its sensitivity to the cost of financial distress and times. Thus, in the current state of knowledge, the *ex ante* approach to the capital structure of non- entrepreneurial companies lacks theoretical foundations.

Summary and Conclusions

In this article I have tried to summarize the results obtained by applying the incomplete contracts approach to corporate governance. In a world where all future observable contingencies can be costlessly contracted upon *ex ante*, there is no room for governance. By contrast, in an incomplete contracts world, corporate governance can be defined as the set of conditions that shapes the *ex post* bargaining over the quasi-rents generated by a firm. A governance system has efficiency effects both *ex ante*, through its impact on the incentive to make relationship-specific investments, and *ex post*, by altering the conditions under which bargaining takes place. A governance system also affects the level and the distribution of risk.

The incomplete contracts approach has been very successful in explaining the corporate governance of entrepreneurial firms. It can explain how ownership is allocated and how capital structure is chosen. By contrast, it is difficult for this approach to cope with the complexity of large publicly traded complies.

Nevertheless, recent contributions in the area are able to account for some important features of large corporations: allocation of ownership to the providers of capital who are dispersed, and the importance of internal organization.

Many aspects, however, remain to be investigated. First and foremost is the role of the board of directors. The second is the interaction between the different mechanisms of corporate governance. While we have many models that describe how each mechanism works in isolation, we know very little about how they interact. The effects are not obvious. For example, debt and takeovers are generally thought, in isolation, to be two instruments that reduce the amount of quasi-rents appropriated by management. But the use of debt may crowd out the effectiveness of takeovers, increasing rather than decreasing managerial rents (Novaes and Zingales 1995). Third, the normative implications of this approach deserve more attention. In a world of incomplete contracts, privately optimal governance can be inefficient *ex ante* and *ex post*. Of course, this is only a theoretical possibility, whose relevance needs to be assessed in the data. The most important contribution, however, will arise from a development of the underlying theory. Without a better understanding of why contracts are incomplete, all the results are merely provisional.

See Also

- ▶ Hold-up Problem
- ▶ Incomplete Contracts

Bibliography

Aghion, P., and P. Bolton. 1992. An incomplete contract approach to financial contracting. *Review of Economic Studies* 59: 473–494.

Aghion, P., P. Bolton, and L. Felli. 1997. *Some issues on contract incompleteness*. Working Paper, London School of Economics.

Alchian, A., and H. Demsetz. 1972. Production, information costs and economic organization. *American Economic Review* 62: 777–795.

Bebchuk, L., and L. Zingales. 1996. *Corporate ownership structures: Private versus social optimality.* Working Paper No. 5584. Cambridge, MA: NBER.

Berle, A., and G. Means. 1932. *The modern corporation and private property*. New York: Macmillan.

Blair, M.M. 1995. *Ownership and control*. Washington, DC: Brookings Institution.

Blair, M.M., and L. Stout. 1997. *A theory of corporation law as a response to contracting problems in team production.* Working Paper, Brookings Institution.

Chandler, A. 1966. *Strategy and structure*. Garden City: Anchor Books.

Fama, F., and M.C. Jensen. 1983a. Separation of ownership and control. *Journal of Law and Economics* 26: 301–325.

Fama, E., and M.C. Jensen. 1983b. Agency problems and residual claims. *Journal of Law and Economics* 26: 327–349.

Grossman, S., and O. Hart. 1980. Takeover bids, the free rider problem and the theory of the corporation. *Bell Journal of Economics* 11: 42–69.

Grossman, S., and O. Hart. 1986. The costs and the benefits of ownership: A theory of vertical and lateral integration. *Journal of Political Economy* 94: 691–719.

Hansmann, H. 1996. *The ownership of enterprise*. Cambridge, MA: Belknap Press of Harvard University Press.

Hart, O. 1989. An economist's perspective on the theory of the firm. *Columbia Law Review* 89: 1757–1774.

Hart, O. 1995. *Firms, contracts, and financial structure*. Oxford: Oxford University Press.

Hart, O., and J. Moore. 1990. Property rights and the nature of the firm. *Journal of Political Economy* 98: 1119–1158.

Hellwig, M. 1997. Unternchmensfinanzierung, Unterichmenskontrolle und Ressourcenallokation: Was leister das Finanzsystem. Arbeitspapier Nr. 97/02, University of Mannheim.

Jensen, M.C., and W. Meckling. 1976. Theory of the firm; managerial behavior, agency costs and ownership structure. *Journal of Financial Economics* 3: 305–360.

Klein, H., R. Crawford, and A. Alchian. 1978. Vertical integration, appropriable rents and the competitive contracting process. *Journal of Law and Economics* 21: 297–326.

La Porta, R., F. Lopez de Silancs, A. Shleifer, and R. Vishny. 1996. *Law and finance.* Working Paper No. 5661. Cambridge, MA: NBER.

La Porta, R., F. Lopez de Silancs, A. Shleifer, and R. Vishny. 1997. Legal determinants of external finance. *Journal of Finance* 52: 1131–1150.

Maskin, F., and J. Tirole. 1997. *Unforeseen contingencies, property rights and incomplete contracts*. Working Paper No. 1796, Institute of Economic Research, Harvard University.

Milgrom, P. 1988. Employment contracts, influence activities, and efficient organization design. *Journal of Political Economy* 96: 42–60.

Milgrom, P., and J. Roberts. 1990. Bargaining costs, influence costs and the organization of economics activity. In *Perspectives on positive political economy*, ed. J. Alt and K. Shepsle. Cambridge: Cambridge University Press.

Myerson, R. 1979. Incentive compatibility and the bargaining problem. *Econometrica* 47: 61–73.

Novaes, W., and L. Zingales. 1995. *Capital structure choice when managers are in control: Entrenchment versus efficiency.* Working Paper No. 5384. Cambridge, MA: NBER.

Rajan, R., and L. Zingales. 1996. *The tyranny of the inefficient: An enquiry into the adverse consequences of power struggles.* Working Paper No. 5396. Cambridge, MA: NBER.

Rajan, R., and L. Zingales. 1998. Power in a theory of the firm. *Quarterly Journal of Economics* 113: 387–432.

Rajan, R., and L. Zingales. 2001. The firm as a dedicated hierarchy. *Quarterly Journal of Economics* 116: 805–851.

Rajan, R., H. Servaes, and L. Zingales. 2000. The cost of diversity: Diversification discount and inefficient investment. *Journal of Finance* 55: 35–80.

Shleifer, A., and R. Vishny. 1989. Management entrenchment: The case of manager- specific investments. *Journal of Financial Economics* 25: 123–140.

Shleifer, A., and R. Vishny. 1997. A survey of corporate governance. *Journal of Finance* 52: 737–783.

Wiggins, S.N., and G.D. Libecap. 1985. Oil field unitization: Contractual failure in the presence of imperfect information. *American Economic Review* 75: 368–385.

Williamson, O. 1985. *The economic institutions of capitalism*. New York: The Free Press.

Zingales, L. 1995a. Insider ownership and the decision to go public. *Review of Economic Studies* 62: 425–448.

Zingales, L. 1995b. What determines the value of corporate votes? *Quarterly Journal of Economics* 110: 1047–1073.

Corporate Law, Economic Analysis of

Robert Daines and Michael Klausner

Abstract

Economic analysis of corporate law largely focuses on (*a*) the efficiency of legal rules and the proper role of the law, (*b*) the ways in which legal rules affect shareholders' ability to monitor managers, and (*c*) the effect of limited liability on the relationship between the corporation and third parties. This article reviews the literature in each of these areas.

Keywords

Agency costs; Collective action problem; Contractarian conception of the corporation; Corporate charters; Corporate control; Ccorporate governance; Corporate law; Corporations; Federalism; Hedge funds; Herding; Insider trading; Law and economics; Learning externalities; Limited liability; Monitoring; Network externalities; Ownership and control; Poison pill; Race to the top/bottom; Shareholder activism; Shareholder suits; State competition; Takeover defence; Takeovers

JEL Classifications

K22; G34; G38

The economic analysis of corporate law focuses primarily on publicly held corporations. Following Coase (1937), the corporation is conceptualized as a nexus of contracts. Because corporate law focuses primarily on the authority of management and its obligations to shareholders, the primary 'contract' of interest is that between management and shareholders. The content of the manager–shareholder contract is conceptualized in terms of the agency-cost model of Jensen and Meckling (1976), with management viewed collectively as agent, and shareholders viewed collectively as principal. Ideally, the terms of the manager–shareholder contract minimize agency costs and thereby maximize the value of the firm.

Most of the economics-oriented corporate law literature can be divided into three areas, all of which focus on the United States. First, there are papers that analyse the economic forces by which corporate law is created by states and adopted by firms, and the proper role of corporate law in light of those forces. Second, there are papers that analyse particular monitoring mechanisms that law creates – shareholder voting, shareholder lawsuits, takeovers. A third set of papers analyses the basic features of a corporation, focusing on limited liability.

This review will discuss these three sets of papers. We do not address the substantial literature on law and finance that suggests that a country's corporate law rules may affect its financial markets and economic growth (see La Porta et al., 1997, 1998; and Rajan and Zingales, 1998). Nor do we address corporate governance strategies, such as CEO pay, that are largely independent of corporate law.

The Role of Corporate Law

Economics-oriented scholarship on the role of corporate law can be roughly divided into two generations. The first generation, which spanned the period from the late 1970s to the mid-1990s, tended to reach the conclusion that market forces would yield socially optimal corporate governance outcomes. The second generation spans the period from the mid-1990s to the present. This generation, which includes more empirical work than the first, has painted a less perfect picture of the relationship between market forces and socially optimal corporate governance (see Klausner, 2006).

First-Generation Scholarship

The central insight of the first generation of economics-oriented corporate law scholarship was the conceptualization of the public corporation as a contractual arrangement between

managers and shareholders. This insight has its origin in Coase (1937). It was developed within the agency cost framework in Jensen and Meckling (1976), and extended to the analysis of corporate law by Easterbrook and Fischel (1989, 1991). Although managers and shareholders do not negotiate governance arrangements, the price mechanism for a company's shares in an initial public offering (IPO) is expected to serve the same function, just as it does in other markets where buyers and sellers do not explicitly negotiate contracts. Consequently, the legally enforceable elements of the corporation's governance structure are viewed as the product of a market-mediated contracting process. Scholars writing in this framework therefore argue that firms' governance structures tend to minimize the agency costs associated with the separation of ownership and control, and thereby maximize the value of the corporation.

Legally enforceable governance commitments can take either of two forms. First, firms select the corporate law rules that govern the rights of shareholders and the obligations of management. Each of the 50 US states has enacted corporate law rules. Firms are free to elect to be governed by any of these rules, regardless of where they do business. To be governed by any state's legal rules, a firm need only incorporate in that state at the time of its IPO. Subsequent disputes between managers and shareholders will then by the decided according to the corporate law of that state. A firm cannot change its state of incorporation unless its board of directors and shareholders holding at least a majority of its shares agree. Second, pre-IPO manager/shareholders must draft a charter that will govern the corporation once it goes public. A charter begins as a blank slate and can include any governance arrangements that a firm's pre-IPO shareholders choose to adopt. To a substantial degree, the law allows a firm's charter to override provisions of corporate law. Thus, corporate law rules are often simply default rules that can be superseded by a corporation's charter terms.

Thus, one insight of this first generation was that corporate law was a product that states produce and firms consume. Winter (1977) was the first to argue that states are engaged in a 'race to the top' to produce corporate law that would tend to minimize agency costs. In order to obtain revenues from franchise fees and to create business for their local lawyers, states were expected to offer corporate law (that is, default rules) that would maximize the value of many firms and thereby save firms the trouble of customizing their own charter terms. Romano (1985) provided empirical evidence consistent with the proposition that a race to the top was occurring. She found, however, that Delaware had already achieved a substantial lead and questioned whether the race would actually make it to the top. The argument that market forces would produce legally enforceable governance commitments that would minimize agency costs stood in contrast to an earlier claim by Cary (1974) that states were engaged in a 'race to the bottom' to create legal rules that appeal to management at the expense of shareholders.

Second-Generation Scholarship

The second generation of scholarship has cast both empirical and theoretical doubt on the contractarian claims described above.

Empirical Findings

A central claim of the contractarian conception of the corporation and corporate law is that corporations are heterogeneous in their corporate governance needs – hence the value of atomistic contracting. Empirical studies have now shown, however, that there is a high degree of uniformity in firms' governance commitments at the time they go public.

Daines (2002) found that, between 1978 and 2002, 50 per cent of firms incorporated in Delaware, and that during the second half of this period over 70 per cent of firms incorporated in Delaware. More importantly, however, Daines found that, among firms that did not incorporate in Delaware, nearly all incorporated in the state in which they were headquartered – whatever that state happened to be. Bebchuk and Cohen (2003) and Kahan (2006) confirmed Daines's findings.

These findings regarding incorporation decisions have three implications. First, the decision to incorporate in one's home state (when no out-of-state firms incorporate there) cannot be motivated primarily by the content of a state's laws. Something else must be at work. Daines's findings suggest that this choice may be made by the firm's local lawyer, hoping to keep the firm's business following the IPO, or by management wanting access to the state legislature if it needs a law passed. Romano (1987) found that most state anti-takeover legislation enacted in the 1980s was initiated by in-state management seeking protection from hostile bids. Bebchuk and Cohen (2003) found that states seem to retain more home-state incorporations if they already have state anti-takeover statutes on their books, but Kahan (2006) refuted this finding. Kahan did find, however, that states with very low-quality corporate law retained fewer home-state incorporations than did other states.

Second, these findings imply a high degree of uniformity in the governance commitments reflected in a firm's incorporation decision. Firms that focus on the quality of corporate law choose Delaware law. This uniformity casts some doubt on the contractarian assumption that firms are heterogeneous in their governance needs. Alternatively, the findings may suggest that there is value in uniformity itself, a point addressed below. Either way, there is evidence that the choice of Delaware as a state of incorporation enhances firm value. Romano (1987) and Daines (2001) found evidence consistent with this conclusion. Subramanian (2004) argues that this is a small-firm effect.

Third, the findings on incorporation choices cast doubt on the proposition that states compete to attract incorporations – whether racing to the top or to the bottom, Delaware seems to be the only state competing. This is what Kahan and Kamar (2002) find in a study of states' efforts, or lack thereof, to attract incorporations and to earn revenues from them.

Empirical research has also revealed a high degree of uniformity in corporate charters. These supposed vehicles of customized contracting and innovation turn out to be fairly empty vessels. The only dimension on which they vary is in that of takeover defences (Klausner, 2006), and variability in that respect sits uneasily with the proposition that IPO charters maximize firm value. Three studies by Daines and Klausner (2001), Field and Karpoff (2002) and Coates (2001) have shown that firms commonly go public with charters providing for staggered boards, which are an effective anti-takeover defence that tends to reduce share value.

Theoretical Challenges to the Contractarian Framework

It is possible that the contractarians overstated their premise that firms are heterogeneous in their governance needs. When it comes to legally enforceable governance commitments, perhaps one size fits all.

There are theoretical reasons, however, to doubt that homogeneous governance needs explain the uniformity described above. The essentially complete absence of customization or innovation in corporate charters suggests there are market imperfections in the contracting process. There has been no lack of innovation in corporate governance since the mid-1980s. None, however, originated in a corporate charter. Innovation at the firm level has taken the form of unilateral adoption of governance structures – for instance, an independent board or separation of CEO and chair – with no legally binding commitment to maintain those structures. The absence of legally binding commitments suggests that the cost of legal enforcement plays some role in the relative emptiness of corporate charters. While there have been innovations in legally enforceable governance mechanisms, they have not occurred at the level of the individual firm charter or even state law, as the contractarian thesis predicts. Instead, they have occurred, for better or worse, through Securities and Exchange Commission (SEC) regulation and federal statute (Sarbanes–Oxley Act, described below).

Klausner (1995, 2006) and Kahan and Klausner (1996) posit that there are learning and network externalities associated with state corporate law and corporate charter terms. As a result of these externalities, commonly used governance

mechanisms have value independent of their intrinsic content; they tend to be better understood and less uncertain in their application than customized mechanisms. These externalities may thus explain the attraction of Delaware and the lack of customization or innovation in corporate charters.

In this context learning externalities take the form of judicial precedents interpreting and applying legal rules, and lawyers' familiarity with these precedents. Because many firms have been incorporated in Delaware, there is a large body of Delaware precedent. As a result, there is less uncertainty regarding how a legal rule will be applied. This may make Delaware valuable because firms have adopted it in the past.

Future judicial interpretations are valuable as well. The larger the number of firms that use the same legal rule or charter term over time, the more the rule or term will be litigated in the future, and the more frequently it will be interpreted. As Hansmann (2006) explains, the alternative would be periodic charter amendments, which could be difficult to accomplish because of the need to have a majority of shareholders and the company's board agree. Consequently, the market dynamic by which firms choose a state of incorporation can be expected to mirror that of product markets in network industries. The equilibrium in those industries can be socially suboptimal uniformity, which may be what is reflected in the attraction of Delaware incorporation and the 'plain vanilla' charter – that is, a charter with no customization that adopts essentially all default rules.

Kahan and Klausner (1996) offer two additional explanations of uniformity in charter terms and incorporation choices. One is that lawyers who draft charters on behalf of their corporate clients may be exhibit the same sort of individually rational herd behaviour that Scharfstein and Stein (1990) and Zwiebel (1995) model for agents such as money managers. These models are based on reputational payoffs to winning or losing with or without the herd. The second explanation relies on results in psychological experiments that reveal a 'status quo' bias, an 'anchoring' bias and a 'conformity' bias in other settings.

Law-Intensive Monitoring Mechanisms

Corporate law creates three monitoring mechanisms and influences a fourth. First, corporate law gives shareholders the right to vote for the board of directors and to approve certain major changes, such as a change to the firm's charter or a merger or sale of the firm. Second, corporate law specifies managers' duties to shareholders and provides a way for shareholders to collectively sue management for its failure to fulfil these obligations. Third, corporate law regulates the takeover process, which allows a poorly run firm to be acquired by a third party. Finally, US federal securities law imposes mandatory disclosure obligations on publicly held firms, which facilitates each of these monitoring mechanisms and enables non-legal monitoring mechanisms (such as the press).

Shareholder Voting

Corporate law gives control of the firm to the board of directors. Shareholder influence over managers comes from their right to elect the board and their implicit (or explicit) threat to vote out incumbent directors. Board elections are held annually and shareholders frequently have the ability to call interim elections. Today, voting is also the means by which control over firms changes hands in a takeover (Gilson and Schwartz, 2001).

Shareholders' ability to oust directors is thus an important check on managerial misbehaviour. The primary limitation on the effectiveness of the shareholder vote is economic rather than legal: shareholders' collective action problems. Individual shareholders with small stakes may not find it worthwhile to become informed and therefore typically either fail to vote or simply vote with management.

An important question is whether institutional investors, by virtue of their larger stakes, will solve the collective action problem and monitor managers more effectively. Money managers, pension funds, mutual funds, banks, insurance companies, and hedge funds all aggregate large pools of equity capital and may be more effective monitors. Rock (1991) and Romano (1993) give

some reason to be cautious about their impact, however. They point out that the interests of money managers sometimes conflict with those of other shareholders. Banks, pension funds and insurance companies may side with incumbent managers if doing so gives them other opportunities to profit by managing the firm's pension funds, making loans or selling other services. Index funds have different disincentives to monitor. They compete on cost, and activism would increase their costs. Public pension funds are frequently active in pressuring managers, but these funds are run by political appointees and may favour politically popular proposals unrelated to firm value. Thus, the empirical evidence suggests that institutional shareholder activism has had only weak effects on firm performance (see, for example, Romano, 2001).

Others, focusing on the rules that govern shareholder ownership and voting, are also cautious about the potential impact about institutional investors. Roe (1994) and Black (1992) argue that shareholder passivity and collective action problems are created not solely by economic forces but also by politically motivated legal constraints that limit the institutional shareholder's incentives and ability to check incumbent managers. In this political view of shareholder passivity, a variety of banking, insurance and financial regulations prevent institutional investors from owning larger stakes or from monitoring managers more closely.

More recently, hedge funds have begun to aggregate large blocks of stock and to use their voting power to influence firm policies. Some investigate whether hedge funds have interests that conflict with other shareholders, which would suggest that hedge fund activism should be regulated (see Kahan and Rock, 2007; and Hu and Black, 2006). The alternative view is that hedge funds' large stakes and relative freedom from regulatory restrictions allows them to overcome collective action problems and to monitor managers.

Shareholder Suits

The law provides mechanisms by which shareholders can collectively sue managers for mismanagement. As a means of controlling agency costs, however, shareholder suits are flawed. Because most shareholders will gain little from a successful lawsuit, shareholders often have no incentive to initiate or monitor these suits. Unless a major institutional shareholder is involved as lead plaintiff, lawyers initiate the suits, pay all costs, make litigation decisions, including settlement decisions, and collect a fee if the plaintiff class collects. To the extent the lawyer's interests diverge from the interests of the shareholders, agency costs are present on the plaintiffs' side of these lawsuits.

On the defendants' side, the familiar agency costs are present. Managers can use corporate funds to protect themselves – appropriately in some cases and inappropriately in others. They use corporate funds to purchase directors' and officers' liability insurance, which covers their personal liability and defence costs, unless they are proven to have engaged in deliberate fraud or the equivalent. Management can also use corporate funds to settle suits. Alexander (1991), Macey and Miller (1991), Coffee (1985, 1986), Romano (1991), Bohn and Choi (1996), among others, have argued that meritorious suits against management settle too easily, and that the prospect of settlement encourages frivolous suits.

The result of this battle of agents is nearly always a settlement in which the corporation and/or its directors' and officers' liability (D&O) insurer are the sole sources of payments. Consequently, payments go from shareholder to shareholder either directly or via insurance companies through premiums. Unless these suits deter mismanagement, the net winners are the lawyers on both sides. The shareholders in the aggregate are net losers (see Arlen and Carney, 1992; Langevoort, 1996; Mahoney, 1992; Easterbrook and Fischel, 1985).

Without commenting on the merits of these suits, Black, Cheffins and Klausner (2006) found only 13 cases, out of several thousand filed since 1980, in which outside directors have made personal payments. Inside managers bear personal liability more often, but settlement dynamics leave their assets untouched in all but a handful of cases per year (Alexander, 1991).

Consequently, there is a question whether these suits have a significant deterrent effect.

The Public Securities Litigation Reform Act of 1995 (PSLRA) created several mechanisms designed to deter the filing of non-meritorious suits and to deter early settlement of meritorious suits. For instance, the law empowered the courts to select a lead plaintiff to monitor the shareholders' lawyer, with a presumption favouring institutional shareholders with substantial shareholdings. The law also requires a court to dismiss a suit unless the plaintiffs have alleged particular facts that support a 'strong inference' that a violation of the securities laws was committed with the legally required intent. This requirement was directed at the reported practice by which lawyers would file suits simply because a company's shares took a sharp drop in price, and then force the company into an expensive discovery process.

Ever since its enactment, scholars have tried to assess the impact of the PSLRA on securities class actions. Event studies, on the whole, have indicated that the law had a positive impact on share prices (Spiess and Tkac, 1997; Johnson et al. 2000a; Johnson et al. 2000b). However, Ali and Kallapur (2001) found that the legislation had a negative impact on share prices. Studies have also tended to show that the PSLRA reduced the filing of non-meritorious suits (Johnson et al. 2000b; Bajaj et al. 2003). Others suggest, however, that some meritorious suits are deterred as well (Choi, 2007; Sale, 1998).

Choi et al. (2005), Thomas and Cox (2006) and Perino (2006) have shown that, while private institutional shareholders have not assumed the role of lead plaintiff, public pension plans have assumed that role to some extent. Perino (2006) found evidence consistent with monitoring by public pension plans.

Market for Corporate Control

The market for corporate control in the United States is regulated by state and federal law and is an important check on agency costs. If a firm is run poorly, either because managers are inattentive, consume too many perks or miss profitable merger opportunities, it may become the target of a takeover and its managers replaced. An active market in corporate control thus gives managers incentive to increase firm value (Manne, 1965).

In a 'hostile takeover' a buyer attempts to purchase a large block of stock, use its voting power to oust incumbent managers, purchase the remaining shares, and replace management. Alternatively, in the shadow of a hostile takeover, managers can agree to a 'friendly merger'. Both are associated with large gains to target shareholders. The evidence generally suggests that the premium comes from improvements in firm performance (see Andrade et al. 2001; Romano, 1992).

The law and economics literature has focused on three questions. First, what should managers do when the firm becomes the target of a hostile takeover? Easterbrook and Fischel (1982, 1991) argue that target management should remain passive and that the law should prohibit them from resisting a takeover. They argue that resistance will reduce bidder returns and thus bidders' incentive to engage in takeovers. This will in turn reduce the disciplinary threat of takeovers and increase agency costs generally. Gilson (1982) and Bebchuk (1982) argue that managers should resist a takeover attempt to the extent necessary to hold an auction, which will assure that the assets of the firm end up in their highest valued use.

A second question involves whether managers' negotiating over a potential merger should be allowed to grant termination fees or 'lock-ups' to favoured bidders. Such measures may discourage competition and affect the outcome of an auction, raising the risk that managers will favour particular bidders in exchange for private benefits, such as job security. Ayres (1990) and Fraidin and Hanson (1994) argue that termination fees and lock-ups will often not change the outcome of the auction and should therefore not be disfavoured. Kahan and Klausner (1996) examine how termination fees and lock-ups affect bidders' incentives to make a bid in the first place and their impact on agency costs generally. They explain that there is no reason for a target to grant a termination fee greater than a bidder's cost of making a bid.

Finally, a large literature examines whether, on average, takeover defences help or harm

shareholder wealth. The typical research strategy examines how a firm's stock price reacts to the adoption of a takeover defence (see, for example, Comment and Schwert, 1995). This strategy usually suffers from a fatal flaw: it ignores the fact that the most potent defence, the 'poison pill', is freely available to all firms even after a hostile bid is received. Therefore, in effect, all firms have a poison pill and most other takeover defences are relatively unimportant. To disable a poison pill, a hostile bidder must first wage a proxy fight to unseat incumbent managers, install new managers who can remove the poison pill, and then go forward with the merger. The only takeover defences that are relevant other than a poison pill are those that either prevent an acquirer from replacing a target board or delay an acquirer's effort to do so. The most common defence is a classified (or staggered) board, which prevents an acquirer from taking control of a target board for two annual election cycles (see, for example, Daines, 2006; Faleye, 2007; Coates and Subramanian, 2002). Dual class stock, which is rarely used, allows management to control the election of the board and can therefore prevent an acquisition altogether.

A related literature examines whether firm takeover defences and shareholder rights predict stock returns (see, for example, Gompers et al. 2003; Cremers and Nair, 2005).

Mandatory Disclosure

The monitoring mechanisms described above all depend, in part, on informed shareholders. Shareholder monitoring (of the kind contemplated by voting, law suits and the market for corporate control) is more effective when investors are informed. Thus, in many ways, the central regulatory event in US financial history was probably the 1933 and 1934 Acts, which created the Securities and Exchange Commission and required that publicly traded firms disclose detailed information about their historical performance and financial condition. These rules force firms both to disclose what they would otherwise prefer to keep private and to keep private information they might otherwise wish to disclose.

It is easy to see why disclosure might be valuable to investors. Accurate information allows investors to price securities and to monitor managers' performance. It is less easy to see why disclosure rules must be mandatory. Firms that fail to disclose will find it hard to raise money, as investors may take silence for bad news and refuse to invest (Ross, 1979; Grossman, 1981; Milgrom, 1981). Therefore, firms and entrepreneurs may find it in their own interest to disclose information, whether or not it is required by law.

However, firms would not always find full disclosure to be in their interest. Disclosure imposes direct costs as firms produce and verify the information, as well as indirect costs if competitors, customers and others can use the information to the firm's disadvantage. Moreover, the costs and benefits of disclosure are likely to vary between firms. Left to their own devices, therefore, firms will commit to varying levels of disclosure. Some therefore argue that markets can sort out the costs and benefits of disclosure and believe that uniform and mandatory disclosure requirements are unnecessary and even harmful (Romano, 2005; Mahoney, 1997; Choi and Guzman, 1997). Others believe that there are externalities from a firm's disclosures and that a mandatory rule may therefore be socially beneficial (Easterbrook and Fischel, 1991; Coffee, 1984; Dye, 1990; Admati and Pfleiderer, 2000).

Empirical evidence has not conclusively resolved this debate. Stigler (1964), Benston and Cohen (1969); Benston (1973) and Simon (1989), report evidence that mandatory disclosure did not improve investor welfare, but may have changed the characteristics of firms that go public. Recent evidence examines the effect of mandatory disclosure on firm returns and on asymmetric information (see Greenstone et al. 2006; Daines and Jones, 2007).

A related debate involves whether managers should be allowed to trade on non-public information. Some hold that trading by informed insiders reveals valuable information and reduces agency costs (Manne, 1966; Carlton and Fischel, 1983). Others argue that insider trading is inefficient (Cox, 1986; Kraakman, 1991) or reduces stock market liquidity (Goshen and

Parchomovsky, 2000). Beny (2007) reviews international evidence.

Creditors and the Corporation

Because the corporation is a legal entity, distinct from its shareholders and managers, shareholders in the firm have 'limited liability' in that they are generally not personally liable for the debts of the corporation. At worst, public shareholders can lose their equity in the firm if the firm becomes insolvent.

This separate legal status gives rise to two issues. First, because shareholders will reap the upside of the firm's successes but will not bear the full downside of its failures, managers may promote the interests of shareholders at the expense of creditors (Jensen and Meckling, 1976). The legal rule of 'veil piercing' developed to respond to this problem, though to a very limited extent. Under extreme circumstances in which a corporation is undercapitalized and other conditions are met, a court may impose liability on the corporation's shareholders. As a practical matter, however, this rule is not applied to public companies' shareholders, and in the private company context the courts' application of this rule is notoriously unpredictable (Thompson, 1991).

The rule of limited liability makes sense for contract creditors, who can negotiate their own protection from default or charge and interest rate that compensates for the risk. Tort creditors, however, are different. Those owed compensation for, say, a firm's pollution emissions, will not have had the opportunity to negotiate with the firm *ex ante* to address the possibility that it will not have sufficient net assets to pay them. Thus, to deter corporate management from externalizing costs in the form of accidents and other torts and to prevent excess investment in risky activities, Hansmann and Kraakman (1991) argue that it may be desirable, and practical, to hold public shareholders personally liable for a corporation's torts. Grundfest (1992) and Alexander (1992) disagree as to the practicality of this proposal.

A second issue involving limited liability is the use of the corporate form to 'partition' assets to create separate pools of assets to bond separate debts and other contractual commitments. Hansmann and Kraakman (2000) explain how the partitioning of assets to separately bond the commitments of the corporate entity, individual shareholders and corporate entities within a group of affiliated corporations can promote efficiencies in creditor monitoring.

Sarbanes–Oxley Act of 2002

The Sarbanes–Oxley Act of 2002 (SOX) introduced sweeping corporate governance mandates on firms whose shares trade on US securities exchanges. Until this legislation, legal rules regarding substantive corporate governance were the province of US state law, and federal law was limited primarily to disclosure requirements. SOX imposed a series of federal requirements on the board operation and structure. Event studies of various legislative events leading to the enactment of SOX yielded mixed results. Li et al. (2004), Jain and Rezaee (2006), and Chhaochhaira and Grinstein (2004) show a positive reaction, but Zhang (2005) shows a negative reaction. Litvak (2007) finds a negative reaction by comparing foreign cross-listed firms subject to SOX with cross-listed firm not subject to SOX. Aggarwal and Williamson (2006) found that six of the governance structures mandated by SOX (all related to board independence) had a positive impact on share value when adopted by firms voluntarily prior to SOX. Romano (2005), on the other hand, looked at other SOX requirements (loans to officers, executive certification of financials, auditors' provision of non-audit services, and audit committee independence) and reports that there is no evidence to support their value to shareholders. Linck et al. (2006) find that whatever the benefit of SOX, it increased the cost of boards, especially for small firms.

See Also

► Corporate governance

Bibliography

Admati, A.R., and P. Pfleiderer. 2000. Forcing firms to talk: Disclosure regulation and externalities. *Review of Financial Studies* 13: 479–519.

Aggarwal, R. and R. Williamson. 2006. Did new regulations target the relevant corporate governance attributes? Working paper, McDonough School of Business, Georgetown University.

Alexander, J.C. 1991. Do the merits matter? A study of settlements in securities class actions. *Stanford Law Review* 43: 497–528.

Alexander, J.C. 1992. Unlimited shareholder liability through a procedural lens. *Harvard Law Review* 106: 387–445.

Ali, A., and S. Kallapur. 2001. Securities price consequences of the Private Securities Litigation Reform Act of 1995 and related events. *Accounting Review* 76: 431–460.

Andrade, G., M. Mitchell, and E. Stafford. 2001. New evidence and perspectives on mergers. *Journal of Economic Perspectives* 15(2): 103–120.

Arlen, J., and W. Carney. 1992. Vicarious liability for fraud on securities markets. *University of Illinois Law Review* 1992: 691–745.

Ayres, I. 1990. Analyzing stock lock-ups: do target treasury sales foreclose or facilitate takeover options? *Columbia Law Review* 90: 682–718.

Bajaj, M., S. Muzumdar, and A. Sarin. 2003. Securities class action settlements: An empirical analysis. *Santa Clara Law Review* 43: 1001–1033.

Bebchuk, L. 1982. The case for facilitating competing tender offers. *Stanford Law Review* 35: 24–47.

Bebchuk, L.A., and A. Cohen. 2003. Firms' decisions where to incorporate. *Journal of Law and Economics* 46: 383–425.

Benston, G.J. 1973. Required disclosure and the stock market: An evaluation of the Securities Exchange Act of 1934. *American Economic Review* 63: 132–155.

Benston, G.J., and A. Cohen. 1969. The value of the SEC's accounting disclosure requirements. *Accounting Review* 44: 515–532.

Beny, L. 2007. Insider trading laws and stock markets around the world. *Journal of Corporate Law* 32: 237–300.

Black, B. 1992. Next steps in proxy reform. *Journal of Corporation Law* 18: 1–55.

Black, B., B. Cheffins, and M. Klausner. 2006. Outside director liability. *Stanford Law Review* 58: 1055–1060.

Bohn, J., and S.J. Choi. 1996. Fraud in the new-issues market: Empirical evidence on securities class actions. *University of Pennsylvania Law Review* 144: 903–964.

Carlton, D., and D. Fischel. 1983. The regulation of insider trading. *Stanford Law Review* 35: 857–895.

Cary, W. 1974. Federalism and corporate law: Reflections upon Delaware. *Yale Law Journal* 83: 663–705.

Chhaochhaira, V., and Y. Grinstein. 2004. Corporate governance and firm value: The impact of the 2002 governance rules. Working paper, Cornell University.

Choi, S.J. 2004. The evidence on securities class actions. *Vanderbilt Law Review* 57: 1465–1525.

Choi, S. 2007. Do the merits matter less after the Private Securities Litigation Reform Act? *Journal of Law Economics and Organization* 23: 598–626.

Choi, S., and A. Guzman. 1997. National laws, international money: Regulation in a global capital market. *Fordham Law Review* 65: 1855–1908.

Choi, S.J., J.E. Fisch, and A.C. Pritchard. 2005. Do institutions matter? The impact of the Lead Plaintiff Provision of the Private Securities Litigation Reform Act. *Washington University Law Quarterly* 83: 869–905.

Coase, R. 1937. The nature of the firm. *Economica* 4: 386–405.

Coates, J.C. IV. 2001. Explaining variation in takeover defenses: Blame the lawyers. *California Law Review* 89: 1301–1415.

Coates, J.C. IV, and G. Subramanian. 2002. The powerful antitakeover force of staggered boards. *Stanford Law Review* 54: 887–951.

Coffee, J.C. Jr. 1984. Market failure and the economic case for a mandatory disclosure system. *Virginia Law Review* 70: 717–753.

Coffee, J.C. Jr. 1985. The unfaithful champion: The plaintiff as monitor in shareholder litigation. *Law and Contemporary Problems* 48: 5–81.

Coffee, J.C. Jr. 1986. Understanding the plaintiff's attorney: The implications of economic theory for private enforcement of law through class and derivative actions. *Columbia Law Review* 86: 669–727.

Comment, R., and G.W. Schwert. 1995. Poison or placebo? Evidence on the deterrence and wealth effects of modern antitakeover measures. *Journal of Financial Economics* 39: 3–43.

Cox, J.D. 1986. Insider trading regulation and the production of information: Theory and evidence. *Washington University Law Quarterly* 64: 475–505.

Cremers, M., and V.B. Nair. 2005. Governance mechanisms and equity prices. *Journal of Finance* 60: 2859–2894.

Daines, R. 2001. Does Delaware law improve firm value? *Journal of Finance and Economics* 62: 525–558.

Daines, R. 2002. The incorporation choices of IPO firms. *New York University Law Review* 77: 1559–1605.

Daines, R. 2006. Do classified boards affect firm value? Takeover defenses after the poison pill. Working paper, Stanford Law School.

Daines, R., and C. Jones. 2007. Mandatory disclosure, asymmetric information and liquidity: The impact of the 1934 Act. Working paper, Law School, Stanford University.

Daines, R., and M. Klausner. 2001. Do IPO charters maximize firm value? Antitakeover protection in IPOs. *Journal of Law Economics and Organization* 17: 83–120.

Dye, R.A. 1990. Mandatory v. voluntary disclosures: The cases of financial and real externalities. *Accounting Review* 65: 1–24.

Easterbrook, F., and D. Fischel. 1982. Auctions and sunk costs in tender offers. *Stanford Law Review* 35: 1–19.

Easterbrook, F., and D. Fischel. 1985. Optimal damages in securities cases. *University of Chicago Law Review* 52: 611–642.

Easterbrook, F., and D. Fischel. 1989. The corporate contract. *Columbia Law Review* 89: 1416–1448.

Easterbrook, F., and D. Fischel. 1991. *The economic structure of corporate law*. Cambridge, MA: Harvard University Press.

Faleye, O. 2007. Classified boards, firm value, managerial entrenchment. *Journal of Financial Economics* 83: 501–529.

Field, L.C., and J.M. Karpoff. 2002. Takeover defenses of IPO firms. *Journal of Finance* 57: 1857–1889.

Fraidin, S., and J. Hanson. 1994. Toward unlocking lockups. *Yale Law Journal* 103: 1739–1834.

Gilson, R. 1982. Seeking competitive bids versus pure passivity in tender offer defense. *Stanford Law Review* 35: 51–67.

Gilson, R.J., and A. Schwartz. 2001. Sales and elections as methods for transferring corporate control. *Theoretical Inquiries in Law* 2: 783–814.

Gompers, P.A., J.L. Ishii, and A. Metrick. 2003. Corporate governance and equity prices. *Quarterly Journal of Economics* 118: 107–155.

Goshen, G. and G. Parchomovsky. 2000. On insider trading, markets, and 'negative' property rights in information. Fordham Law and Economics Research Paper No. 06.

Greenstone, M., P. Oyer, and A. Vissing-Jorgensen. 2006. Mandated disclosure, stock returns, and the 1964 Securities Acts Amendments. *Quarterly Journal of Economics* 121: 399–460.

Grossman, S.J. 1981. The information role of warranties and private disclosure about product quality. *Journal of Law and Economics* 24: 461–483.

Grundfest, J.A. 1992. The limited future of unlimited liability: A capital markets perspective. *Yale Law Review* 102: 387–425.

Hansmann, H. 2006. Corporation and contract. *American Law and Economics Review* 8: 1–19.

Hansmann, H., and R. Kraakman. 1991. Toward unlimited shareholder liability for corporate torts. *Yale Law Review* 100: 1879–1934.

Hansmann, H., and R. Kraakman. 2000. The essential role of organizational law. *Yale Law Journal* 110: 387–440.

Hu, H., and B. Black. 2006. The new vote buying: Empty voting and hidden ownership. *Southern California Law Review* 79: 811–908.

Jain, P., and Z.J. Rezaee. 2006. The Sarbanes–Oxley Act of 2002 and capital-market behavior: Early evidence. *Contemporary Accounting Research* 23: 629–654.

Jensen, M.C., and W.H. Meckling. 1976. Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics* 3: 303–360.

Johnson, M., R. Kasanik, and K. Nelson. 2000a. Shareholder wealth effects of the Private Securities Litigation Reform Act of 1995. *Review of Accounting Studies* 5: 217–233.

Johnson, M., K. Nelson, and A. Pritchard. 2000b. In re Silicon Graphics Securities Litigation: Shareholder wealth effects of the interpretation of the Private Securities Litigation Reform Act's pleading standard. *Southern California Law Review* 73: 773–809.

Johnson, M., R. Kasznik, and K. Nelson. 2001. The impact of securities litigation reform on the disclosure of forward-looking information. *Journal of Accounting Research* 39: 297–328.

Kahan, M. 2006. The demand for corporate law: Statutory flexibility, judicial quality, or takeover protection. *Journal of Law, Economics, & Organization* 22: 340–365.

Kahan, M., and E. Kamar. 2002. The myth of state competition in corporate law. *Stanford Law Review* 55: 679–760.

Kahan, M., and M. Klausner. 1996. Path dependence in corporate contracting: Increasing returns, herd behavior and cognitive biases. *Washington University Law Quarterly* 74: 347–366.

Kahan, M., and M. Klausner. 1997. Standardization and innovation in corporate contracting (or 'The economics of boilerplate'). *Virginia Law Review* 83: 713–755.

Kahan, M., and E. Rock. 2007. Hedge funds in corporate governance and corporate control. *University of Pennsylvania Law Review* (forthcoming).

Klausner, M. 1995. Corporations, corporate law, and networks of contracts. *Virginia Law Review* 81: 757–833.

Klausner, M. 2006. The contractarian theory of corporate law: A generation later. *Journal of Corporate Law* 31: 779–797.

Kraakman, R. 1991. The legal theory of insider trading regulation in the United States. In *European insider dealing*, ed. K. Hopt and E. Wymeerisch. London: Butterworths.

La Porta, R., F. Lopez-de-Silanes, A. Shleifer, and R.W. Vishny. 1997. Legal determinants of external finance. *Journal of Finance* 52: 1131–1150.

La Porta, R., F. Lopez-de-Silanes, A. Shleifer, and R.W. Vishny. 1998. Law and finance. *Journal of Political Economy* 106: 1113–1155.

Langevoort, D.C. 1996. Capping damages for open-market securities fraud. *Arizona Law Review* 38: 639–668.

Li, H., Pincus, M. and Rego, S. 2004. Market reaction to events surrounding the Sarbanes–Oxley Act of 2002 and earnings management. Working paper, University of Iowa.

Linck, J., Netter, J. and Yang, T. 2006. Effects and unintended consequences of the Sarbanes–Oxley Act on corporate boards. Working paper, University of Georgia.

Litvak, K. 2007. The effect of the Sarbanes–Oxley Act on non-US companies cross-listed in the US. *Journal of Corporate Finance.* 13: 195–228.

Macey, J., and G. Miller. 1991. The plaintiffs' attorney's role in class action and derivative litigation: Economic analysis and recommendations for reform. *Chicago Law Review* 58: 1–94.

Mahoney, P. 1992. Precaution costs and the law of fraud in impersonal markets. *Virginia Law Review* 78: 623–660.
Mahoney, P.G. 1997. The exchange as regulator. *Virginia Law Review* 83: 1453–1500.
Manne, H. 1965. Mergers and the market for corporate control. *Journal of Political Economy* 73: 110–120.
Manne, H. 1966. *Insider trading and the stock market*. New York: Free Press.
Milgrom, P. 1981. Good news and bad news: Representation theorems and application. *Bell Journal of Economics* 12: 380–391.
Perino, M. 2006. Institutional activism through litigation: An empirical analysis of public pension fund participation in securities class actions. Legal Studies Research Paper No. 06–0055, St. John's University.
Pritchard, A.C., and H. Sale. 2005. What counts as fraud? An empirical study of motions to dismiss under the Private Securities Litigation Reform Act. *Journal of Empirical Legal Studies* 2: 125–149.
Rajan, R.G., and L. Zingales. 1998. Financial dependence and growth. *American Economic Review* 88: 559–586.
Rock, E. 1991. The logic and (uncertain) significance of institutional shareholder activism. *Georgetown Law Review* 79: 445–506.
Roe, M. 1994. *Strong managers, weak owners*. Princeton: Princeton University Press.
Romano, R. 1985. Law as a product: Some pieces of the incorporation puzzle. *Journal of Law Economics and Organization* 1: 225–283.
Romano, R. 1987. The political economy of takeover statutes. *Virginia Law Review* 73: 111–198.
Romano, R. 1991. The shareholder suit: Litigation without foundation? *Journal of Law Economics and Organization* 7: 55–87.
Romano, R. 1992. A guide to takeovers: Theory, evidence and regulation. *Yale Journal on Regulation* 9: 119–180.
Romano, R. 1993. Public pension fund activism in corporate governance reconsidered. *Columbia Law Review* 93: 795–852.
Romano, R. 1998. Empowering investors: A market approach to securities regulation. *Yale Law Journal* 107: 2359–2430.
Romano, R. 2001. Less is more: Making institutional investor activism a valuable mechanism of corporate governance. *Yale Journal on Regulation* 18: 174–252.
Romano, R. 2005. The Sarbanes–Oxley Act and the making of quack corporate Governance. *Yale Law Journal* 114: 1521–1611.
Ross, S.A. 1979. Disclosure regulation in financial markets: Implications of modern finance theory and signaling theory. In *Issues in financial regulation: Regulation of American Business and Industry*, ed. F.R. Edwards. New York: McGraw Hill.
Sale, H. 1998. Heightened pleading and discovery stays: An analysis of the effect of the PSLRA's Internal-Information Standard on '33 and '34 Act claims. *Washington University Law Quarterly* 76: 537–595.
Scharfstein, D.S., and J. Stein. 1990. Herd behavior and investment. *American Economic Review* 80: 465–489.
Schwert, G.W. 1995. Comment: Poison or placebo: Evidence on the deterrence and wealth effects of modern antitakeover measures. *Journal of Financial Economics* 39: 3–43.
Simon, C. 1989. The effect of the 1933 Securities Act on investor information and the performance of new issues. *American Economic Review* 79: 295–318.
Spiess, D.K., and P. Tkac. 1997. The Private Securities Litigation Reform Act of 1955: The stock market casts its vote. *Managerial and Decision Economics* 18: 545–561.
Stigler, G. 1964. Public regulation of the securities markets. *Journal of Business* 2: 117–142.
Subramanian, G. 2004. The disappearing Delaware effect. *Journal of Law Economics and Organization* 20: 32–59.
Thomas, R., and J. Cox. 2006. Does the plaintiff matter? An empirical analysis of lead plaintiffs in securities class actions. *Columbia Law Review* 100: 101–155.
Thompson, R. 1991. Piercing the corporate veil: An empirical study. *Cornell Law Review* 76: 1036–1074.
Thompson, R., and H. Sale. 2003. Securities fraud as corporate governance: Reflections upon federalism. *Vanderbilt Law Review* 56: 859–915.
Thompson, R., and R. Thomas. 2004. The public and private faces of derivative lawsuits. *Vanderbilt Law Review* 58: 1747–1823.
Winter, R. 1977. State law, shareholder protection and the theory of the corporation. *Journal of Legal Studies* 6: 251–292.
Zhang, I. 2005. Economic consequences of the Sarbanes–Oxley Act of 2002. Related Publication 05–07. AEI–Brookings Joint Center for Regulatory Studies.
Zwiebel, J. 1995. Corporate conservatism, herd behavior and relative compensation. *Journal of Political Economy* 103: 1–25.

Corporations

Randall Morck

Abstract

A corporation is an artificial person with many of the rights of a biological one. The first business corporations pooled the savings of many individuals to permit ventures on a scale none could afford individually. Most large American and British corporations lack controlling shareholders; the consequent lack of monitoring and

control gives rise to corporate governance problems reflecting the private benefits of control. The view that corporations should be run to maximize shareholder conflicts in many countries with the actual legal duties of corporate officers, and collides with evidence that stock prices are sometimes set by investors with incomplete information.

Keywords

Agency costs; Corporate governance; Corporations; Limited liability; Monitoring; Noise traders; Private benefits of control; Stock markets; Time inconsistency

JEL Classifications

M1

A *corporation* is an artificial person, with many of the legal rights of a biological one. This modern legal and economic usage arose in the 16th century from the term's now archaic meaning of 'a group acting as one body' – encompassing municipal governments, businesses and other groups of individuals united towards a common goal. In that century and the next, trade with the Orient and the New World promised immense returns, but only after vast capital outlays on fleets of ships, networks of forts and private armies to defend them. The first business corporations, such as the Dutch East Indies Company, the British East India Company and the Hudson's Bay Company, were formed to pool the savings of many individuals and permit ventures on a scale none could afford individually. Each owner of a *share* of the corporation was periodically entitled to a *dividend* – a pro rata division of the corporation's free cash flow.

Polling all a corporation's shareholders for each business decision was impractical in an age of sailing ships and horse-drawn carriages. Instead, the shareholders elected *boards of governors* (later *directors*) – reputable men trusted by the majority of shareholders to direct the corporation's affairs.

This did not prevent all dispute. The Dutch East Indies Company (*Vereenigde Oostindische Compagnie* in Dutch) was formed as a limited time venture. When that limit drew near, the board boldly announced that the corporation would persist indefinitely. The shareholders sued to force a liquidating dividend – and lost! Fortunately, they found they could sell their shares to other investors for the value of a liquidating dividend – or even more (Frentrop 2002/3). Thus was born the first modern stock market, and the *alienability,* or unhindered sale, of shares became a defining characteristic of a corporation. Letting shareholders realize their investments by selling their shares, rather than liquidating the business, gave corporations a second defining characteristic: *indefinitely long lives.*

Boards occasionally betrayed their shareholders' trust and caused a corporation to contravene the law. Since individual shareholders were not consulted, holding them fully to account for the corporation's misdeeds seemed wrong. Since the corporation is a legal person, plaintiffs could sue it directly, and need not sue its shareholders personally. Thus, *limited liability* statutes came to shield individual shareholders from personal lawsuits for wrongs by corporations whose shares they own. Limited liability, a third defining characteristic of the modern business corporation, is an important innovation because it frees individuals to invest their savings in corporations run by strangers, undertaking risky ventures, or doing business in far off places. Vulnerability to personal lawsuits would otherwise make such investments seem indefensibly reckless to most savers.

Early corporations, like the Hudson's Bay Company, assigned one vote to each share in board elections. This essentially let the wealthiest shareholders appoint the board and, if they wished, run the corporation in their narrow interest rather than in the interests of all shareholders equally. For example, a large shareholder might force the corporation to do business with another corporation she controlled on disadvantageous terms. This sort of self-dealing, which (Johnson et al. 2000) dub 'tunneling', remains a widespread corporate governance concern where firms typically have dominant shareholders. Or a dominant shareholder might simply relish the perks, power and prestige of running the corporation, and

refuse to make way for more qualified managers – a corporate governance problem called 'entrenchment' (Morck et al. 1988). Entrenchment and tunnelling provide controlling shareholders with *private benefits of control* – returns not shared with small shareholders (Dyck and Zingales 2004). Distorted corporate governance associated with private benefits of control remains a first-order governance concern wherever corporations typically have a controlling shareholder. According to (La Porta et al. 1999), this includes the large corporate sectors of virtually all countries except Germany, Japan, the United Kingdom and the United States. Small and middle-sized corporations everywhere tend to have controlling shareholders.

In the 19th century, *democratic* corporate governance became associated with *one vote per shareholder,* rather than one vote per share (Dunlavy 2004). Echoes of this remain in the *voting caps* of modern Canadian and European corporations, which limit any single shareholder's voting power regardless of shares owned. However, large shareholders in many countries later turned deviations from one vote per share to their advantage by granting themselves special classes of common stock with many votes per share. In most countries, such *dual class shares* now virtually always magnify, rather than limit, the voting power of large shareholders, and so amplify, rather than dampen, problems associated with private benefits of control (Nenova 2003).

In the United States and the United Kingdom, however, one vote per share is the norm in shareholder meetings. Disclosure rules, regulatory oversight, officer and director liability, and other restraints on private benefits of control also seem more effective in America and Britain than elsewhere in curtailing private benefits of control (La Porta et al. 1999; Dyck and Zingales 2004). This makes being a large shareholder less attractive, especially if holding a diversified portfolio of small stakes in many firms reduces risk (Burkart et al. 2003). Unsurprisingly, most large American and British corporations now lack controlling shareholders (La Porta et al. 1999). They are run by professional managers who often own few shares (Morck et al. 1988).

A small shareholder who monitored and controlled these corporate top managers would bear all the investigative, legal and administrative costs involved, but the benefits of better governance would be spread across all shareholders. The cost therefore typically exceeds the benefit for any small shareholder acting alone (Grossman and Hart 1988). The consequent general lack of monitoring and control in corporations with no large shareholder gives rise to *other people's money* corporate governance problems. Adam Smith (1776) famously explains that, since corporate managers who own few or no shares are more

> the managers of other people's money than of their own, it cannot well be expected that they should watch over it with the same anxious vigilance with which partners in a private copartnery frequently watch over their own. Like the stewards of a rich man, they .. . . consider attention to small matters as not for their master's honour and very easily give themselves a dispensation from having it.

Unmonitored professional managers can thus enjoy the perks and privileges of running large corporations without any real concern for the returns they generate. Berle and Means (1932) argue that this sort of governance problem occurs in many large American corporations.

But in other countries, other people's money governance problems probably also afflict many corporations that, on first inspection, seem to have a controlling shareholder. This is because large corporations in most countries are not freestanding entities, but belong to *corporate groups* (La Porta et al. 1999). These are typically pyramidal structures, in which an apex shareholder, usually an extremely wealthy family, controls one or more listed corporations, which each control more listed corporations, which each control yet more listed corporations, *ad valorem et infinitum.* A family that controls 51 per cent of a listed corporation that controls 51 per cent of another that controls 51 per cent of yet another and so on actually owns only 0.51^n of the corporation n tiers down the in pyramid, with the remainder of each corporation financed by public or minority shareholders. Pyramids with a dozen or more layers are not uncommon, rendering the controlling shareholder's actual ownership of corporations at the

pyramid's base negligible. Pyramidal business groups thus permit controlling shareholders to extract private benefits of control from corporate empires financed largely from other people's money (Morck et al. 2000; Bebchuk et al. 2000). Pyramids were common in the United States until the 1930s (Berle and Means 1932; Bonbright and Means 1932), but were eliminated by various New Deal initiatives, including the double and multiple taxation of inter-corporate dividends (Morck 2005). British pyramids apparently withered under sustained attacks from institutional investors (Franks et al. 2005). However, the relevant unit of economic analysis for many purposes elsewhere in the world should often be the *business group,* not the corporation.

Jensen and Meckling (1976) show that *agency costs,* the present value of the costs of expected future governance shortfalls of any sort, are born by the corporation's initial shareholders. A corporation's founders receive less per share when they first sell shares to outside investors if worse corporate governance problems seem likely.

This gives rise to a *time inconsistency* problem in securities and corporations law. Investors and entrepreneurs selling shares to the public benefit from credible guarantees of good governance because these limit agency costs and so raise share prices. But top corporate decision makers in firms that have already issued shares, who foresee issuing no more, wish to maximize their utility (Baumol 1959, 1962; Williamson 1964) and understandably value the freedom to spend public shareholders' money as they like and to capture such private benefits of control as they can. Actual public policy probably reflects these groups' relative political lobbying power, which can change over time (Morck et al. 2000, 2005).

The normative view that a corporation *should* be run to maximize shareholder value derives from economists' assumption that firms maximize profits. In neoclassical economic theory, a firm that maximizes the present value of all its expected future economic profits necessarily maximizes the market value of its shares. This follows from modelling the corporation as a *nexus* of contracts, with the shareholders the *residual claimants* to the firm's cash flows (Fama and Jensen 1983a, b). Neoclassical theory further allows that profit maximization (value maximization in a multi-period setting) accords with economic efficiency under certain idealized conditions; see, for example, (Varian 1992) and (Malliaris and Brock 1983).

This normative view conflicts with the actual legal duties of corporate officers, directors and controlling shareholders in many countries. For example, many northern European countries and some US states impose a duty to balance shareholders' interests with those of *stakeholders,* especially employees. This is formalized in the German legal principle of *Mitbestimmung* (co-determination), which requires members of the *Aufsichtsrat* (supervisory board) of a large corporation to balance the interests of shareholders, employees and the state (Fohlin 2005). Common law legal systems assign officers and directors a duty to act *for the corporation.* In Britain and the United States, this is often interpreted as a duty to act for the corporation's owners, its shareholders. A duty to maximize share value seems implicit (Jensen and Meckling 1976; Black and Coffee 1997). However, the Canadian Supreme Court holds in *Peoples v. Wise* that the duty of the officers and directors of a corporation is not to shareholders, nor to any other stakeholders, but to the corporation per se. The social welfare implications of assigning different legal duties to corporate top decision makers are incompletely understood. Giving labour a voice in corporate decision making seems to impede risk taking and hamper growth (Faleye et al. 2006). Moreover, regardless of their assigned objective, if those entrusted to govern great corporations occasionally put their own interests ahead of their legal duties, agency costs must arise in some form.

The view that a corporation's top managers ought to maximize shareholder value also collides with evidence that stock prices are sometimes set by investors with incomplete information (Myers and Majluf 1984) or behavioural biases (Shleifer 2000). Coase (1937) argues that firms come about to alleviate information asymmetries and other

market imperfections, collectively denoted *transactions costs,* and that the boundaries of the firm correspond to an efficient solution to these problems. (Alchian and Demsetz 1972) argue that the critical market imperfections arise from people working in teams. (Williamson 1975) argues that interdependent assets are more generally important. Jensen (2004) calls for more research on normative theories about the boundaries of the corporation and the objective function of its top decision makers if stock prices are set by *noise traders,* that is, investors with behavioural biases. One approach holds that corporations actually exist primarily to lock the economy's capital into productive uses by isolating capital allocation decisions from maniac or panicked investors (Stout 2004). This view long dominated discussions of corporate management in Japan (for example, Aoki and Dore 1994) but appears to give rise to its own set of inefficiencies (see, for example, Morck and Nakamura 1999).

See Also

- ▶ Corporate Governance
- ▶ Firm, Theory of the

Bibliography

Alchian, A., and H. Demsetz. 1972. Production, information costs and economic organization. *American Economic Review* 62: 777–795.

Aoki, M., and R.P. Dore. 1994. *The Japanese firm: The sources of competitive strength*. New York: Oxford University Press.

Baumol, W. 1959. *Business behavior, value and growth*. New York: Macmillan.

Baumol, W. 1962. On the theory of expansion of the firm. *American Economic Review* 52: 1078–1087.

Bebchuk, L., R. Kraakman, and G. Triantis. 2000. Stock pyramids, cross ownership and dual class equity: The mechanisms and agency costs of separating control from cash flow rights. In *Concentrated corporate ownership*, ed. R. Morck. Chicago: University of Chicago Press.

Berle, A., and G. Means. 1932. *The modern corporation and private property*. New York: Macmillan.

Black, B.S., and J.C. Coffee Jr. 1997. Hail Britannia? Institutional investor behavior under limited regulation. *Michigan Law Review* 92: 1997–2087.

Bonbright, J., and G. Means. *1932. The holding company – Its public significance and its regulation*. New York: McGraw-Hill.

Burkart, M., F. Panunzi, and A. Shleifer. 2003. Family firms. *Journal of Finance* 58: 2173–2207.

Coase, R. 1937. The nature of the firm. *Economica* 4: 386–405.

Dunlavy, C. 2004. *The unnatural origins of one vote per share – A chapter in the history of corporate governance*. Working paper, Department of History, University of Wisconsin, Madison.

Dyck, A., and L. Zingales. 2004. Private benefits of control: An international comparison. *Journal of Finance* 59: 537–601.

Faleye, O., V. Mehrotra, and R. Morck. 2006. When labor has a voice in corporate governance. *Journal of Financial and Quantitative Analysis* 41: 489–510.

Fama, E., and M. Jensen. 1983a. Agency problems and residual claims. *Journal of Law and Economics* 26: 327–349.

Fama, E., and M. Jensen. 1983b. Separation of ownership and control. *Journal of Law and Economics* 26: 301–325.

Fohlin, C. 2005. The history of corporate ownership and control in Germany. In *The history of corporate governance around the world: Family business groups to professional managers*, ed. R. Morck. Chicago: University of Chicago Press.

Franks, J., C. Mayer, and S. Rossi. 2005. Spending less time with the family: The decline of family ownership in the UK. In *The history of corporate governance around the world: Family business groups to professional managers*, ed. R. Morck. Chicago: University of Chicago Press.

Frentrop, P. 2002/3. *A history of corporate governance*. Amsterdam: Deminor Press.

Grossman, S., and O. Hart. 1988. One share one vote and the market for corporate control. *Journal of Financial Economics* 20: 175–202.

Jensen, M. 2004. Agency costs of overvalued equity. Harvard NOM research paper no. 04-26. Harvard Business School.

Jensen, M., and W. Meckling. 1976. Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics* 3: 305–360.

Johnson, S., R. La Porta, F. Lopez-de-Silanes, and A. Shleifer. 2000. Tunneling. *American Economic Review* 90: 22–27.

La Porta, R., F. Lopez-de-Silanes, and A. Shleifer. 1999. Corporate ownership around the world. *Journal of Finance* 54: 471–517.

Malliaris, A.G., and W. Brock. 1983. *Stochastic methods in economics and finance*. Amsterdam: North-Holland.

Morck, R. 2005. How to eliminate pyramidal business groups: The double-taxation of intercorporate dividends and other incisive uses of tax policy. *Tax Policy and the Economy* 19: 135–179.

Morck, R., and M. Nakamura. 1999. Banks and corporate control in Japan. *Journal of Finance* 54: 319–340.

Morck, R., A. Shleifer, and R. Vishny. 1988. Management ownership and market valuation: An empirical analysis. *Journal of Financial Economics* 20: 293–315.
Morck, R., D.A. Stangeland, and B. Yeung. 2000. Inherited wealth, corporate control, and economic growth: The Canadian disease. In *Concentrated corporate ownership*, ed. R. Morck. Chicago: University of Chicago Press.
Morck, R., D. Wolfenzon, and B. Yeung. 2005. Corporate governance, economic entrenchment, and growth. *Journal of Economic Literature* 43: 655–720.
Myers, S., and N. Majluf. 1984. Corporate financing and investment decisions when firms have information that investors do not have. *Journal of Financial Economics* 13: 187–222.
Nenova, Tatiana. 2003. The value of corporate voting rights and control: A cross-country analysis. *Journal of Financial Economics* 68: 325–351.
Shleifer, A. 2000. *Inefficient markets: An introduction to behavioral finance*. Oxford: Oxford University Press.
Smith, Adam. 1776. *An inquiry into the nature and causes of the wealth of nations*. London: Ward, Lock, and Tyler.
Stout, Lynn. 2004. On the nature of corporations. Law & economics research paper no. 04-13. School of Law, UCLA.
Varian, H. 1992. *Microeconomic analysis*. 3rd ed. New York: W. W. Norton.
Williamson, O. 1964. *The economics of discretionary behavior: Managerial objectives in a theory of the firm*. Englewood Cliffs: Prentice Hall.
Williamson, O. 1975. *Markets and hierarchies: Analysis and antitrust implications*. New York: Free Press.

Corporatism

Joseph Halevi

Corporatism is a set of political doctrines aimed at organizing civil society on the basis of professional and occupational representation in chambers called Estates or Corporations. It maintains that class conflict is not inherent in the capitalist system of production and ownership relations. Corporatism has its ideological roots mainly in 19th-century French and Italian Catholic social thought, as well as in German romanticism and idealism. Corporative ideas can be found in eminent European thinkers. Hegel, in his *Philosophy of Right*, thought of a corporate structure in which the Estates constituted the link between civil society and the State (Hegel 1821). In France, Durkheim put forward a view of corporatism specifically related to the division of labour engendered by modern industry. According to Durkheim, the Corporations' task is to diversify at the level of each industry the general principles of industrial legislation formulated by the political assemblies (Durkheim 1893).

The Catholic strand appeared first as a response to the social cleavages stemming from the industrialization of Europe. It advocated a return to the corporate form of guild associations of the Middle Ages, which it romantically viewed as based on social harmony. In 1891 the papal encyclical Rerum Novarum took a more reformist approach. It rejected the notion that 'class is naturally hostile to class, and that the wealthy and the working men are intended by nature to live in mutual conflict' (Rerum Novarum 1891; in Camp 1969, p. 81). At the same time it recognized the legitimacy of independent worker's unions, although preference was given to the creation of a single organization embracing employers and employees. In practice the Catholic movement opted for the first variant, partly because the industrialists rejected the idea of a single organization and partly because of the strength of the Socialist-led unions.

Where politics were concerned, in countries like Italy and Germany, the Catholics gradually reconciled their corporative social views with parliamentarism. In other instances, the Catholic movement aimed at supplanting parliamentary institutions altogether. In Austria, for example, the alliance between the Social Christians and the fascist Heimwehr was the basis of the corporative Constitution passed before the assassination of Chancellor Dolfuss, in 1934.

Germany produced an important theoretician of corporatism: Karl Marlo. He wrote a comprehensive critique of liberalism in favour of Estate organizations (Marlo 1885). His views are a reaction to the radicalization of the working class, which led to the 1848 Revolution. In that year, Marlo proposed to the Frankfurt Parliament that it form a social chamber composed by the representatives of all occupations whose task would be to formulate the social legislation to be approved by the political chamber.

Modern corporatism begins with the idealist jurist and Italian nationalist Alfredo Rocco. In his conception corporatism was an instrument for fostering the productive power of the nation. He considered the Estates to be merely organs of the State.

Italian fascism absorbed Rocco's views from its inception, although it combined them with elements of Catholic corporatism as well as with aspects of the doctrine of revolutionary syndicalism held by Georges Sorel (Togliatti 1970). The syndicalist component was eliminated in 1926 when Rocco, who had become Mussolini's Minister of Justice, legally recognized the fascist unions only, banning all the others in existence. Under the pressure of the employers' association, Italian Confederation of Industry, shop floor committees, which the syndicalists wanted to retain, were also outlawed. The Italian corporative state was institutionalized when in 1927 a labour charter (Carta del Lavoro) was promulgated and, in 1934, a law was issued establishing 22 Estates. In 1939 their 500 delegates formed the Camera dei Fasci e delle Corporazioni, which replaced the Chamber of Deputies.

Italy's corporative state did not coordinate economic activity. Instead, it enabled the Government to control labour relations by making tutelage over the newly created labour unions legal. It enforced arbitration tribunals formed by a judge and two experts, thereby excluding any kind of worker representation even from the fascist unions (Salvemini 1936; Rossi 1955).

The rescue operations to save the Banca Commerciale which led to the formation in 1933 of the state-holding IRI (Institute for Industrial Reconstruction) are to be linked to the impact of the Depression on the endemic banking crisis in Italy rather than to any corporative economic programme. Already in 1922, Piero Sraffa pointed out that the frequent crises of Italy's banking system were caused by the fact that bank's activities were based on lending short while borrowing long. Sraffa showed that this was a structural characteristic of the Italian economy (Sraffa 1922). The Depression magnified the above tendencies and the Government found itself compelled to intervene on an unprecedented scale.

The corporative juridical structure only played an indirect economic role. It legalized, as part of the Estates, a very subordinate form of unionism, while allowing the employers to struggle – within the Estates – for the creation of Consortia which, once approved, become compulsory (Rossi 1955). Here there is both a similarity and a difference vis-à-vis the German case. The National Socialist regime pursued a policy of forced cartelization – an objective shared by many industrial groups well before 1933 – but not through a legal system of a syndicalist, corporative character. Workers were organized in a completely separate body called the Labour Front (Neumann 1944; Kuczynski 1945).

The juridically more complete nature of Italian corporatism became a reference for populist movements in South America. One important example is the Estado Novo established in Brazil under President Getulio Vargas in the years 1937–46. Following the Italian pattern a Labour Charter was issued. The decree-laws of 1939 legalized government prerogatives over labour unions, which were exercised by the Ministry of Labour.

Unlike Italy, Brazilian corporatism allowed the emergence of strong reformist demands. Although labour relations were governed by norms which prevented the formation of alliances between different groups of workers, the process leading to the corporative state marked also the appearance of formal unionism. Hence in Brazil during the liberal phase (1946–64) populist forces were capable of using institutions designed to control the working class for the purpose of giving political power to labour leaders (Erickson 1977). Yet the strengthening of corporatism came from the conservative forces themselves, which after the coup d'état of 1964 tightened the controls over labour organizations.

The main element of modern corporatism consists in a detailed network of technical and juridical norms, enforced by ministerial bodies, aimed at controlling the labour movement. A formal system of Estates had either an incidental character (Italy) or was never implemented.

The economic views of the main advocates of corporatism have never reached an analytical dimension. During the 1930s in Italy some discussion took place around the issue of home corporativus versus *homo oeconomicus* (Mancini et al. 1982).

See Also

- ▶ Economic Theory of the State
- ▶ Fascism

References

Camp, R. 1969. *The papal ideology of social reform. A study in historical development, 1878–1967*. Leiden: E.J. Brill.

Durkheim, E. 1893. *The division of labour in society*. London: Macmillan, 1933.

Erickson, K. 1977. *The Brazilian Corporate State and working class politics*. Berkeley: University of California Press.

Hegel, G. 1821. *Hegel's philosophy of right*. London: Oxford University Press, 1967.

Kuczynski, J. 1945. *Germany: Economic and labour conditions under fascism*. New York: International Publishers.

Mancini, O., F. Parillo, and E. Zagari. 1982. *La teoria economica del corporativismo*. Naples: Edizioni Scientifiche Italiane.

Marlo, K. 1885. *Untersuchungen über die Organisation der Arbeit*. Tübingen.

Neumann, F. 1944. *Behemoth; the structure and practice of national socialism, 1933–1944*. New York: Octagon Books.

Rossi, E. 1955. *Padroni del vapore e fascismo*. Bari: Laterza.

Salvemini, G. 1936. *Under the axe of fascism*. New York: Viking Press.

Sraffa, P. 1922. The bank crisis in Italy. *Economic Journal* 32: 178–197.

Togliatti, P. 1970. *Lectures on fascism*. New York: International Publishers, 1976.

Correspondence Principle

Federico Echenique

Keywords

Comparative statics; Correspondence principle; Monotone models; Strategic complementarities; Tâtonnement; Walrasian general equilibrium

JEL Classifications

D5

The correspondence principle is the relation, which exists in certain economic models, between comparative statics of equilibria and the properties of out-of-equilibrium dynamics.

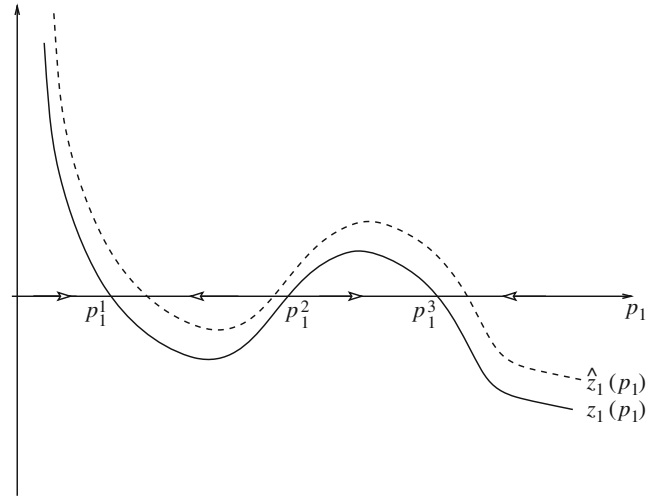
The correspondence principle (CP) implies that one obtains unambiguous comparative statics by selecting equilibria with desirable dynamic properties. Generally, the CP determines comparative statics in models with a one-dimensional endogenous variable, and in monotone multidimensional models. It does not determine comparative statics in general multidimensional models, such as Walrasian general equilibrium models with more than two goods.

One-Dimensional Models

The CP holds quite generally in one-dimensional models. Consider, for example, a two-good economy with excess-demand function for good 1 given by z_1, shown in Fig. 1. We fix the price of good 2; by Walras's Law the equilibrium prices are the zeroes of z_1: there are three equilibria, p_1^1, p_1^2 and p_1^3.

Now consider the comparative-statics exercise of shifting excess demand up to $\hat{z}_1$. What is the effect on equilibrium price? Locally, the price increases if the equilibrium is p_1^1 or p_1^3, but it decreases if it is p_1^2. The different comparative statics at p_1^1 and p_1^2 corresponds exactly to the different behavior of tâtonnement dynamics after a small perturbation: p_1^1 is stable while p_1^2 is unstable.

The difference between comparative statics at p_1^1 and at p_1^2 is easy to explain. The comparative statics at p_1^1 says: slightly larger prices than p_1^1 are reached by increasing excess demand, and smaller prices are reached by decreasing excess demand. Since excess demand is zero at p_1^1, there must be *positive* excess demand at slightly larger prices and *negative* excess demand at slightly smaller prices. Hence, tâtonnement dynamics, which respond to the sign of excess demand, converges to p_1^1 after a small perturbation from p_1^1. On the other hand, at p_1^2 larger prices result from a decrease in excess demand; hence excess demand is positive at larger prices. Similarly, excess



Correspondence Principle, Fig. 1 Two-good economy

demand is negative at smaller prices. As a result, tâtonnement dynamics will not approach p_1^2 after a small perturbation from p_1^2.

If the economy is subject to sporadic shocks, one should not observe p_1^2, the unstable equilibrium. Hence, as a consequence of the correspondence between comparative statics and dynamics, one should expect an increase in excess demand to produce an increase in equilibrium price.

I shall give a general statement of the correspondence principle for the one-dimensional case. Consider a model where the endogenous variable takes values in [0, 1] and equilibria are determined as the fixed points of $f(\cdot, t)$: $[0, 1] \to [0, 1]$; $t \in T \subseteq R$ is an exogenous parameter. Assume that T is convex and that f is C^1.

A *selection of equilibria* is a function $e : T \to [0,1]$ such that $e(t) = f(e(t), t)$ for all $t \in T$. Say that a fixed point $x \in [0, 1]$ is *stable* if there is a neighbourhood V of x such that any sequence x_n satisfying $x_0 \in V$ and $x_{n+1} = f(x_n)$ for $n \geq 1$, converges to x. Say that $x \in [0, 1]$ is *unstable* if, for any neighbourhood V of x, there is a neighbourhood W of x such that all sequences defined as above eventually lie in the complement of W.

Proposition 1 *Let f be monotone increasing in t. If e is a continuous selection of equilibria that is strictly decreasing over some interval $[\underline{t}, \overline{t}]$, then for all $t \in (\underline{t}, \overline{t})$, $e(t)$ is unstable.*

Multidimensional Models

The one-dimensional CP is a relation between the sign of the comparative-statics change in prices, and the sign of excess demand for smaller and larger prices. When more than one price is determined, this relation does not need to exist. Still, the CP holds for monotone models – models where the different dimensions of the endogenous variables are in some sense complements. Monotone economic models stem mainly from game theoretic models with strategic complementarities.

I proceed to give a statement of the CP. Consider a model where the endogenous variable takes values in a compact rectangle $X \subseteq R^n$, and equilibria are determined as the fixed points of $f(\cdot, t) : X \to X; t \in T \subseteq R$ is a parameter and T is convex.

Proposition 2 *Let f be monotone increasing in (x, t) and let e be a continuous selection of equilibria.*

- *If e is strictly decreasing over $[\underline{t}, \overline{t}] \subseteq T$, then for all $t \in (\underline{t}, \overline{t})$, $e(t)$, is unstable.*

- *If e is strictly increasing over $[\underline{t}, \overline{t}]$, then for all $t \in (\underline{t}, \overline{t})$, if $e(t)$ is locally isolated, it is stable.*

Literature

The CP was formulated by Paul Samuelson (1941, 1942, 1947), who also coined the term (though Hicks 1939, stated the CP informally). Samuelson formulated the onedimensional CP. The version in Proposition 1 is taken from Echenique (2000).

Bassett et al. (1968) study the scope of the CP. Arrow and Hahn (1971) present a critical discussion of the CP, and, because it fails in economies with more than two goods, conclude that 'very few useful propositions are derivable from this principle'. The monotone multidimensional CP is from Echenique (2002), who presents a general version of Proposition 2. Echenique (2004) presents a CP that does not rely on continuous selections of equilibria. The CP is also effective in dynamic optimization models (Brock 1983; Burmeister and Long 1977; Magill and Sheinkman 1979) and in models of international trade (Bhagwati et al. 1987).

See Also

► Comparative Statics

Bibliography

Arrow, K.J., and F.H. Hahn. 1971. *General competitive analysis*. San Francisco: Holden-Day.

Bassett, L., J. Maybee, and J. Quirk. 1968. Qualitative economics and the scope of the correspondence principle. *Econometrica* 36: 544–563.

Bhagwati, J.N., R.A. Brecher, and T. Hatta. 1987. The global correspondence principle: A generalization. *American Economic Review* 77: 124–132.

Brock, W.A. 1983. A revised version of Samuelson's correspondence principle. In *Models of economic dynamics*, ed. H. Sonnenschein. New York: Springer-Verlag.

Burmeister, E., and N.V. Long. 1977. On some unresolved qsts in capital theory: An application of Samuelson's correspondence principle. *Quarterly Journal of Economics* 91: 289–314.

Echenique, F. 2000. Comparative statics by adaptive dynamics and the correspondence principle. Working paper no. E00-273, Department of Economics, University of California, Berkeley.

Echenique, F. 2002. Comparative statics by adaptive dynamics and the correspondence principle. *Econometrica* 70: 833–844.

Echenique, F. 2004. A weak correspondence principle for models with complementarities. *Journal of Mathematical Economics* 40: 145–152.

Hicks, J.R. 1939. *Value and capital*. Oxford: Clarendon Press.

Magill, M.J.P., and J.A. Sheinkman. 1979. Stability of regular equilibria and the correspondence principle for symmetric variational problems. *International Economic Review* 20: 297–315.

Samuelson, P.A. 1941. The stability of equilibrium: Comparative statics and dynamics. *Econometrica* 9: 97–120.

Samuelson, P.A. 1942. The stability of equilibrium: Linear and nonlinear systems. *Econometrica* 10: 1–25.

Samuelson, P.A. 1947. *Foundations of economic analysis*. Cambridge, MA: Harvard University Press.

Correspondences

M. Ali Khan

Abstract

Correspondences are versatile mathematical objects for which a rich theory can be developed. They arise naturally in many diverse areas of applied mathematics, including economic theory. For example, an individual consumer's demand correspondence associates with each price system the set of utility maximizing consumption plans. Similarly, an individual producer's supply correspondence associates with each price system the set of profit-maximizing production plans. These individual responses are correspondences rather than functions because of the constancy of marginal rates of substitution in consumption and in production over a range of commodity bundles.

Keywords

Berge's maximum theorem; Brouwer's fixed point theorem; Correspondences; Functions; Kakutani's fixed point theorem; Lyapunov's theorem

JEL Classifications
C0

A correspondence Q from a domain set X to a range set Y associates with each element x in X, a non-empty subset of Y, $Q(x)$. A *function* is a correspondence such that $Q(x)$ is a singleton for each x in X. It is for this reason that a correspondence is also termed a *multi-valued function* or, more simply, a *multi-function*. Another name for a correspondence is a *set-valued mapping*.

Correspondences arise naturally in economic theory. One may think of an individual consumer's demand correspondence, which associates with each price system the set of utility maximizing consumption plans; see, for example, Hildenbrand (1974, p. 92). An equally pervasive example is an individual producer's supply correspondence which associates with each price system the set of profit- maximizing production plans (see, for example, Arrow and Hahn 1971, pp. 54–5). The fact that these individual responses are correspondences rather than functions is simply a consequence of 'flats' in the underlying indifference surfaces and isoquants or, more precisely, of the constancy of marginal rates of substitution in consumption and in production over a range of commodity bundles. Indeed, the association of these marginal rates with the point at which they are evaluated is another example of a correspondence that arises naturally in economic theory, particularly in the study of marginal cost pricing equilibria in economies with increasing returns to scale (for example, Brown et al. 1986). The fact that there is no unique rate of substitution is simply a consequence of 'kinks' in the underlying function. In the case of a convex function, such a correspondence is termed the *subdifferential* correspondence, and, for more general functions, it is *Clarke's generalized derivative*.

If the domain and range of a correspondence are *topological spaces*, one can formulate various notions of continuity of a correspondence. Recall that (X, τ_X) is a topological space if X is a set and τ_X is a collection of subsets of X that contains X and the empty set ø and is closed under finite intersection and arbitrary union. We can now present one formalization of the intuitive idea of continuity of a correspondence. A correspondence $Q : X \to Y$, X, Y, both topological spaces, is said to be upper *semicontinuous* (u.s.c.) if for any V in τ_Y, the set $\{x \in X : Q(x) \subset V\}$ is in τ_X. Q is said to be lower *semicontinuous* (l.s.c.) if for any V in τ_Y, the set $\{x \in X : Q(x) \cap V \neq \text{ø}\}$ is in τ_X. It is easy to convince oneself that a correspondence may be u.s.c. without being l.s.c. and vice versa. It is also easy to show that, if Y is a compact space, a correspondence Q is u.s.c. if and only if its graph, $\text{Gr}Q$, $\text{Gr}Q = \{(x, y) \in X \times Y : y \in Q(x)\}$, is such that its complement belongs to $\tau_X \times \tau_Y$. A correspondence is said to be *continuous* if it is both u.s.c. and l.s.c.

A very useful result for establishing u.s.c. of correspondences arising from maximization is Berge's *maximum theorem*. This states, in particular, that for any continuous correspondence Q from a topological space X to a topological space Y and any continuous function f from $X \times Y$ into the reals, the associated correspondence $\mu : X \to Y$ given by $\mu(x) = \{y \in Q(x) : f(y, x) \geq f(y', x)$ for all $y' \in Q(x)\}$ is u.s.c. This theorem is used to show u.s.c. of the demand and supply correspondences in the theory of the consumer and of the producer.

A result which plays a significant role in the proof of the existence of a competitive equilibrium is Kakutani's *fixed point theorem* for convex valued, u.s.c. correspondences which take a non-empty convex compact subset of an Euclidean space to itself. The theorem states that such correspondences Q have a fixed point, that is, an element x such that $x \in Q(x)$. Kakutani's theorem yields as an immediate corollary Brouwer's fixed point theorem and generalizes, word for word, to locally convex spaces as has been shown by Glicksberg and Ky Fan (see, for example, Berge 1963, p. 251).

It is of interest to know of conditions under which a correspondence $Q : X \to Y$ yields a *continuous selection*, that is, a continuous function $f : X \to Y$ such that $f(x) \in Q(x)$ for all x in X. The celebrated selection theorems of Michael (see, for example, Bessaga and Pelczynski 1975, ch. II.7) give a variety of sufficient conditions for this. One of these requires X to be a paracompact topological space, Y to be a separable Banach space and

Q to be convex valued and l.s.c. This theorem has been used by Gale and Mas-Colell (1974) to show the existence of competitive equilibrium for economies in which consumer preferences need neither be complete nor transitive. If Q is u.s.c. rather than l.s.c., recent work of Cellina gives sufficient conditions under which one may obtain an *approximate continuous selection*.

So far in this exposition we have been considering results on correspondences whose domain and range are both topological spaces. An alternative setting is one where the range is a topological space but the domain is a *measurable space*. $(T, \sum)$ is a measurable space if T is a set and Σ is a family of subsets that includes T and is closed under complementation and countable unions, that is, Σ is a σ-*algebra*. Such correspondences arise naturally in the study of economies in which the set of agents is modelled as a measurable space. An obvious example of such a correspondence is one which associates with every agent his/her set of utility maximizing consumption plans under a given price system.

One can develop concepts analogous to continuity for correspondences from a measurable space to a topological space. A correspondence $Q : T \to Y$ is said to be *measurable* if, for any set V in τ_Y, the set $\{t \in T : Q(t) \cap V \neq \varnothing\}$ is an element of Σ. Variants of this definition have been presented in the literature along with conditions under which these variants are all equivalent. One particularly fruitful variant requires the measure space to be *complete* and the correspondence to have a *measurable graph*, that is, GrQ is a subset of $\sum \bigotimes \mathscr{B}(Y)$, the smallest σ-algebra generated by the sets in $\sum \times \mathscr{B}(Y)$ and where $\mathscr{B}(Y)$ is the smallest σ-algebra generated by sets τ_Y.

We can now state a measure-theoretic analogue of Berge's theorem. Let Q be a correspondence with a measurable graph and f a $\sum \bigotimes \mathscr{B}(Y)$ measurable function. From $T \times Y$ into the reals. Then a result due to the collective efforts of Debreu and Castaing–Valadier states that under a mild restriction on Y, namely Souslin, the correspondence $\mu : T \to X$, $\mu(t) = \{x \in Q(t) : f(t, x) \geq f(t, x')$ for all x' in $Q(t)\}$, has a measurable graph.

We have developed enough terminology to state a fundamental theorem due to the collective efforts of von Neumann, Aumann and St. Beuve. This states that under a restriction on the range space Y, namely, Souslin, every correspondence Q with a measurable graph yields a *measurable selection*, that is, a measurable function $f : T \to Y$ such that $f(t) \in Q(t)$ for all t in T.

Once we have a measurable selection theorem, we are in a position to formulate a satisfactory notion of an integral of a correspondence, a notion which may also be seen as a formalization of a sum of an infinite number of sets. However, one preliminary notion that still needs to be stated is that of a *measure* μ on $(T, \sum)$. A measure μ is a set-valued function from Σ into (say) Euclidean space R^n such that

$$\mu(A) \geq 0, \mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i)$$

for all A, A_i in Σ and such that A_i are mutually disjoint. Now let us assume we know how to integrate a function with respect to μ and can therefore specify a function $f : T \to R^n$ to be an *integrable function* if its integral (Lebesgue integral) is finite. Following Aumann, we can define the integral of a correspondence, Q, $\int_T Q(t)\mathrm{d}\mu$ to be the set $\{\int_T f(t)\mathrm{d}t : f$, an integrable function which is a measurable selection from $Q\}$. It is now clear that $\int_T Q(t)\mathrm{d}\mu$ is non-empty if Q has a measurable graph and if there exists an integrable function g with non-negative values and such that $|x| \leq g(t)$ for all $x \in Q(t)$ and for all $t \in T$.

Finally, we can state a consequence of Lyapunov's theorem on the range of an *atomless measure* that has played a fundamental role in the development of the theory of economies with a continuum of agents. A measure μ on a measurable space $(T, \sum)$ is atomless if $(T, \sum, \mu)$ has no *atoms*, that is $A \in \Sigma$ such that $\mu(A) > 0$ and $B \in \sum, B \subset A$ implies $\mu(B) = \mu(A)$ or $\mu(B) = 0$. The Lyapunov–Richter theorem states that the integral of a correspondence $Q : T \to R^n$ is convex if μ is an atomless measure on (T, Σ).

In summary, a correspondence is a versatile mathematical object for which a deep and rich theory can be developed and which arises naturally in many diverse areas of applied

mathematics, including economic theory. For an introduction to this theory and to its applications, the reader is referred to the following references which also contain all the concepts and results not referenced in this entry.

See Also

- ▶ Fixed Point Theorems
- ▶ Lyapunov Functions

Bibliography

Arrow, K.J., and F.H. Hahn. 1971. *General competitive analysis*. San Francisco: Holden-Day.

Aubin, J.P., and A. Cellina. 1984. *Differential inclusions*. New York: Springer-Verlag.

Berge, C. 1963. *Topological spaces*. New York: Macmillan.

Bessaga, C., and A. Pelczynski. 1975. *Selected topics in infinite-dimensional topology*. Warsaw: Polish Scientific Publishers.

Brown, D., G. Heal, M. Ali Khan, and R. Vohra. 1986. On a general existence theorem for marginal cost pricing equilibria. *Journal of Economic Theory* 38: 111–119.

Castaing, C., and M. Valadier. 1977. *Convex analysis and measurable multifunctions*, Lecture notes in mathematics no. 580. New York: Springer-Verlag.

Clarke, F.H. 1983. *Optimization and nonsmooth analysis*. New York: John Wiley.

Gale, D., and A. Mas-Colell. 1974. An equilibrium existence theorem for a general model without ordered preferences. *Journal of Mathematical Economics* 2: 9–15. Erratum in *Journal of Mathematical Economics* 6: 297–298.

Hildenbrand, W. 1974. *Core and equilibria of a large economy*. Princeton: Princeton University Press.

Klein, E., and E.A. Thompson. 1985. *Introduction to the theory of correspondences*. New York: John Wiley.

Rockafellar, R.T. 1970. *Convex analysis*. Princeton: Princeton University Press.

Corruption and Economic Growth

Niloy Bose

Abstract

Theory is divided over the effects of corruption on economic growth. However, the growing consensus based on the empirical literature is that corruption is associated with negative growth outcomes. This relationship is not necessarily linear, and causality between corruption and economic growth can run in both directions.

Keywords

Corruption; Economic growth; Multiple equilibria; Threshold effects; Two-way causality

JEL Classifications

D73; O11; O17

Corruption can take many forms. Broadly defined, it is the use of public office to promote personal gain (Jain 2001). In recent years, various international development agencies have taken an unequivocal stand that corruption is 'the single greatest obstacle to economic and social development' (www.worldbank.org/oublicsector/anticorrupt). The literature on corruption, however, offers two equally plausible yet opposing views.

Two Opposing Theoretical Views

One view argues that corruption can enhance efficiency and raise growth in the presence of cumbersome bureaucratic regulations. Bribes for instance are sometimes accepted in exchange for overcoming institutional rigidities that raise inefficiencies (Leff 1964; Huntington 1968; Leys 1970). More recent expositions of this view can be found in Acemoglou and Verdier (1998, 2000), who argue that some corruption may be optimal in the presence of incomplete contracts or market failures. Others (Lui 1985; Beck and Maher 1986), without relying on pre-existing institutional rigidities, have argued that corruption introduces competition for government resources and helps to provide services more efficiently. Despite finding abundant support in the day-to-day experiences in many developing countries, this efficiency-enhancing or 'grease the wheel' view has been challenged on the grounds that the regulations and red tape that corruption helps to

circumvent are not exogenous, but are put in place to maximize income from corrupt practices in the first place (Myrdal 1968; Bardhan 1997; Rose-Ackerman 1978).

The other view advocates that corruption lowers the volume and efficiency of private and public investment and therefore is detrimental to economic growth. Theory and anecdotal evidence suggest various channels through which these effects could materialize. For example, corruption could lower resources available for productive public investments (Blackburn et al. 2010) and could divert funds to where bribes are easiest to collect, imparting a bias towards low-productivity projects (Wade 1985; Hardin 1993). Others suggest that corruption changes incentives, prices and opportunities in such a way that allocation of talent, technology and capital move away from their socially most productive use. For example, opportunities to seek rent through corrupt practices could divert investment from human to political capital (Murphy et al. 1991; Ehrlich and Lui 1999); rent-seeking by public officials through the imposition of excessive regulation and bureaucracy could discourage innovation (e.g. Murphy et al. 1993) and encourage informal and inefficient sectors (Sarte 2000).

Empirical Evidence

A number of organizations – most notably Business International Corporation, Political Risk Services Inc., and Transparency International – provide cross-country measures of corruption which are constructed on the basis of subjective evaluations of experts, and surveys sent to a network of correspondents around the world. While they differ in various aspects (including their coverage, methodology and availability) and are susceptible to the usual caveats associated with survey data, these measures are highly correlated, suggesting that they do in fact contain relevant information. The advent of these corruption measures has sparked a flurry of empirical investigations into the relationship between corruption and economic growth. The major findings from these investigations are as follows.

First, there is overwhelming evidence from cross-country analysis that corruption hurts economic growth in various ways: by lowering investment (Mauro 1995; Knack and Keefer 1995), creating obstacles to doing business and encouraging unofficial sectors (e.g. Johnson et al. 1997), reducing inflows of foreign capital (Wei 2000) and decreasing the quality of public investment through a misallocation of public expenditure (Tanzi and Davoodi 1997; Mauro 1998). A number of micro studies (see Svensson 2005 for details) have also yielded insights about the long-run cost of corruption. Despite finding a weak direct association between corruption and the growth rate of GDP, these studies have helped solidify the view that corruption hurts growth through its adverse effect on key determinants of growth. In contrast, there is little evidence to support the efficiency-enhancing view even in countries that are reportedly mired with regulations. Using firm-level survey data, Kaufmann and Wei (1999) in fact find that firms that pay more bribes are likely to spend more time with bureaucrats negotiating regulations, and accordingly face a higher cost of capital.

Second, cross-country differences in the incidence of corruption owe much to cross-country differences in the level of prosperity. According to Treisman (2000), a significant proportion – as much as 50 to 73 per cent – of the variation in corruption indices can be explained by the variation in per-capita income. Other studies (Paldam 2002) have confirmed this relationship.

Third, there is evidence that the growth effect of corruption could be nonlinear. Mendez and Sepulveda (2006) and Bose et al. (2008) provide evidence for multiple regimes – one in which the incidence of corruption is high, and its effect on the quality of public infrastructure is strongly negative, and one for relatively low levels of corruption, where its effect is neutral or perhaps even slightly positive. There could be many reasons for such nonlinearities. For example, corruption could have a smaller (negative) impact where it is more 'predictable' (Campos et al. 1999) and where institutional quality is low (Aidt et al. 2008). Alternatively, these nonlinearities could arise through informational frictions in the public procurement process (Bose et al. 2008): when firms earn

economic profits due to market conditions, some of these profits can be extracted in the form of bribes without affecting the provision of public goods. However, when corruption exceeds a certain threshold, this will no longer continue to hold true.

Finally, there is evidence for multiple equilibria in the corruption–economic growth nexus, leading to the notion that some countries may be drawn into a vicious cycle of low growth and high corruption (Bardhan 1997). Haque and Kneller (2009) provide a formal account of such persistence in the data by identifying corruption-development 'clubs' where countries appear to become trapped in highcorruption–low-development or low-corruption–high-development patterns. At micro-level, explanation for persistence of corruption is typically obtained by appealing to the notion of strategic complementarities, where an individual's incentive to be corrupt depends on the behavior of the others (Andvig and Moene 1990; Murphy et al. 1993). These theories are useful in explaining persistence when corruption has already taken a firm grip on a society. From a practical perspective, what one would, however, like to know is how an economy might settle in one equilibrium rather than another as a result of the interplay between the fundamental determinants of corruption and growth. Some progress (Blackburn et al. 2006; Mauro 2004; Blackburn et al. 2010; Aidt et al. 2008) has been made in addressing this concern. Generally, these papers present a framework where a feedback loop from growth to corruption is combined with standard mechanisms through which corruption reduces growth. This gives rise to a self-enforcing dynamic that provides an explanation why corruption and poverty are often not transient phenomena, but an integral part of the fabric of society.

There is now a large body of evidence supporting the view that corruption is detrimental to economic growth. This evidence, however, is based on perceived indices of corruption – that is, data that do not measure corruption itself, but only capture opinions about its prevalence. Recently, researchers have turned their attention to corruption data that are constructed on the basis of actual experience. (For details, see www.transparency.org/policy research/surveysindices/gcb and http://info.worldbank.org/governance/wbes/.) While still in their state of infancy, investigations (Treisman 2007) based on new data have begun to show promise in advancing our understanding of the corruption–growth relationship. At the same time, there is now wider recognition among researchers that corruption is a multifaceted phenomenon with roots in political, cultural and moral aspects of society. A better understanding of how these factors interact with economic fundamentals and shape incentives for corruption could shed further insights into the corruption–growth nexus.

See Also

▶ State Capture and Corruption in Transition Economies

Bibliography

Acemoglou, D., and T. Verdier. 1998. Property rights, corruption and the allocation of talent: A general equilibrium approach. *Economic Journal* 108: 1381–1403.

Acemoglou, D., and T. Verdier. 2000. The choice between market failures and corruption. *American Economic Review* 90: 194–211.

Aidt, T., J. Dutta, and V. Sena. 2008. Governance regime, corruption and growth: Theory and evidence. *Journal of Comparative Economics* 36: 195–220.

Andvig, J.C., and K.O. Moene. 1990. How corruption may corrupt. *Journal of Economic Behavior and Organization* 13: 63–76.

Bardhan, P. 1997. Corruption and development: A review of issues. *Journal of Economic Literature* 35: 1320–1346.

Beck, P.J., and M.W. Maher. 1986. A comparison of bribery and bidding in thin markets. *Economic Letters* 20: 1–5.

Blackburn, K., N. Bose, and M.E. Haque. 2006. The incidence and the persistence of corruption in economic development. *Journal of Economic Dynamics and Control* 30: 2447–2467.

Blackburn, K., N. Bose, and M.E. Haque. 2010. Endogenous corruption in economic development. *Journal of Economic Studies* 37: 4–25.

Bose, N., A. Murshid, and S. Capasso. 2008. Threshold effects of corruption: Theory and evidence. *World Development* 36: 1173–1191.

Campos, J.E., D. Lien, and S. Pradhan. 1999. The impact of corruption on investment: Predictability matters. *World Development* 27: 1059–1067.

Ehrlich, I., and F.T. Lui. 1999. Bureaucratic corruption and endogenous economic growth. *Journal of Political Economy* 107: 270–293.
Haque, M.E., and R. Kneller. 2009. Corruption clubs: Endogenous thresholds in corruption and development. *Economics of Governance* 10: 345–373.
Hardin, B. 1993. *Africa: Dispatches from a fragile continent*. London: Harper Collins.
Huntington, S.P. 1968. *Political order in changing societies*. New Haven: Yale University Press.
Jain, A.K. 2001. Corruption: A review. *Journal of Economic Surveys* 15: 71–121.
Johnson, S.K., D. Kaufmann, and A. Shleifer. 1997. The unofficial economy in transition. *Brooking Papers on Economic Activity* 2: 159–239.
Kaufmann, D., and Wei, S. J. 1999. Does 'grease money' speed up the wheels of commerce? NBER working paper 7093.
Knack, S., and P. Keefer. 1995. Institutions and economic performance: Cross-country tests using alternative institutional measures. *Economics and Politics* 7: 207–227.
Leff, N.H. 1964. Economic development through bureaucratic corruption. *American Behavioral Scientist* 8: 8–14.
Leys, C. 1970. What is the problem about corruption? In *Political corruption: Readings in comparative analysis*, ed. A.J. Heindenheimer. New York: Holt Rinehart.
Lui, F.T. 1985. An equilibrium queuing model of bribery. *Journal of Political Economy* 93: 760–781.
Mauro, P. 1995. Corruption and growth. *Quarterly Journal of Economics* 110: 681–712.
Mauro, P. 1998. Corruption and composition of government expenditure. *Journal of Public Economics* 31: 215–236.
Mauro, P. 2004. The persistence of corruption and slow economic growth. *IMF Staff Papers* 51: 1–17.
Mendez, F., and F. Sepulveda. 2006. Corruption, growth and political regimes: Cross country evidence. *European Journal of Political Economy* 22: 82–98.
Murphy, K.M., A. Shleifer, and R. Vishny. 1991. The allocation of talent: Implications for growth. *Quarterly Journal of Economics* 106: 505–530.
Murphy, K.M., A. Shleifer, and R. Vishny. 1993. Why rent-seeking so costly to growth? *American Economic Review: Papers and Proceedings* 83: 409–414.
Myrdal, G. 1968. *Asian Drama: An enquiry into the poverty of nations*, vol. 2. New York: Random House.
Paldam, M. 2002. The cross-country pattern of corruption: Economics, culture and the seesaw dynamics. *European Journal of Political Economy* 18: 215–240.
Rose-Ackerman, S. 1978. *Corruption: A study in political economy*. London/New York: Academic Press.
Sarte, P.D. 2000. Informality and rent-seeking bureaucracies in a model of long-run growth. *Journal of Monetary Economics* 46: 173–197.
Svensson, J. 2005. Eight questions about corruption. *Journal of Economic Perspectives* 19: 19–22.
Tanzi, V., and Davoodi, H. 1997. Corruption, public investment and growth. IMF Working paper 97/139.
Treisman, D. 2000. The causes of corruption: A cross-national study. *Journal of Public Economics* 76: 399–457.
Treisman, D. 2007. What have we learned about the causes of corruption from ten years of crossnational empirical research? *Annual Review of Political Science* 10: 211–244.
Wade, R. 1985. The market for public office: Why the Indian state is not better at development. *World Development* 13: 467–497.
Wei, S.J. 2000. How taxing is corruption on international investors? *Review of Economics and Statistics* 82: 1–11.

Cossa, Luigi (1831–1896)

G. De Vivo

Born in Milan, Cossa was Professor of Political Economy at the University of Pavia from 1858 to his death. He was influential both through his works and, perhaps even more, through his many pupils: Pantaleoni (who was not one of them) wrote in 1909 (p. 755) that Cossa was one of the 'three men [who] have been the direct teachers of all Italian economists' (the others being F. Ferrara and A. Messedaglia). Cossa is generally regarded as one of the Italian 'Socialists of the Chair', and as such he was attacked by Ferrara, who accused the 'Germanists' of being 'socialists, and corrupters of the Italian youth' (thus Cossa himself summarized Ferrara's onslaught: Cossa 1876, p. 226).

Cossa had studied in Germany with Roscher and had been strongly influenced by him. The German influence is mainly revealed in his acceptance of the idea of the historical relativity of economic laws. He also maintained that a system of protection 'at certain times and under certain conditions... has given notable advantages to industrial organisation and progress' (Cossa 1876, p. 124). (All this, of course, sounded like blasphemy to Ferrara, a great admirer of Bastiat.) He was far from regarding the German economists as faultless, and never denied the importance, for some parts of political economy, of the deductive

method. A good account of Cossa's position is that given by Edgeworth, who described him as 'hold [ing] the balance between the claims of historical observation and deductive reasoning with great fairness' (Edgeworth 1892, p. 685).

Cossa did not go deeply into economic theory, keeping to a rather superficial eclecticism. But Pantaleoni wrote of him that he had the capacity of perfectly understanding books that he would have never been able to write (Pantaleoni 1898, p. 250).

Cossa's fame is mostly due to the bibliographical essays he published, especially his 1876 *Guida allo Studio dell'Economia Politica*, the 2nd edition of which was translated into English, and published by Macmillan, with a preface by Jevons, in 1880 (a new, greatly enlarged, edition issued in 1892, under the title *Introduzione allo Studio dell'Economia Politica* and translated into English, was also very successful). The *Guida*, like most of Cossa's works, mainly consisted of a 'Historical Part', containing an annotated bibliography of political economy. Another book which also attained some fame, and was translated into many languages (including even Japanese, if we can trust Loria 1896, p. 488) was *Primi Elementi di Economia Politica* (1875).

Cossa's bibliographies can still be instructive for a modern reader, especially the parts on Italian, and on French and German, economics. More interesting than those in his books are however those published in many instalments in the *Giornale degli Economisti*, from 1891 to 1900. These have only recently been reprinted in a single volume (L. Cossa, *Saggi bibliografici di economia politica*, con Prefazione di L. Dal Pane 1963), but have not been translated into English. Cossa's scholarship could be exaggerated (Einaudi spoke of him as '*onnisciente*'). For instance, he attributed to Edward Gibbon Wakefield the 1804 *Essay upon Political Economy* (see p. 245 of the volume just quoted) which had been written by his uncle Daniel (in 1804, E.G. Wakefield was only 9 years old). It must however be said that his standards (especially when compared with those of many of his contemporaries) were generally of a fairly high level.

Selected Works

1875. *Primi elementi di economia politica*. Milan: Hoepli.
1876. *Guide to the study of political economy*. Translated from the second Italian edition [1877], with a Preface by W.S. Jevons, F.R.S. London: Macmillan, 1880.
1893. *An introduction to the study of political economy*. Trans. Louis Dyer. London: Macmillan.
1963. *Saggi bibliografici di economia politica*. Bologna: Forni.

Bibliography

Edgeworth, F.Y. 1892. Review of L. Cossa, *Introduzione allo Studio dell' Economia Politica* (1892). *Economic Journal* 2: 685–687.
Loria, A. 1896. Obituary: L. Cossa. *Económic Journal* 6: 488–490.
Pantaleoni, M. 1898. Dei criteri che devono informare la storia delle dottrine economiche. As reprinted in *Erotemi di Economia*, vol. I. Bari: G. Laterza, 1925.
Pantaleoni, M. 1909. Messedaglia, Angelo. In *Dictionary of political economy*, ed. R.H.I. Palgrave. London: Macmillan. Appendix.

Cost and Supply Curves

James C. Moore

In microeconomic theory we usually suppose that an individual firm has a production technology which can be characterized by a production function $\phi : \Re_+^n \to \Re_+$; where the quantity $\phi : (v)$, for $\in \Re_+^n$, is interpreted as the maximum quantity of output which can be produced, given the vector of quantities of inputs, v. Using the generic notation 'x' to denote the quantity of output, we also suppose that the firm's revenue and cost are described by functions $R : \Re_+ \times P \to \Re_+$ and $K : \Re_+^n \times \Omega \to \Re_+$, where:

$R(x, \rho)$ is the revenue obtained by selling output $x \in \Re_+$, given the market conditions for its

output represented by $\rho \in P$; where ρ is assumed to be outside the firm's control, and

P is the space of possible output market conditions,

$K(\upsilon, w)$ is the cost incurred by the firm in employing the vector of input quantities $\upsilon \in \Re^n_+$, given the input market conditions $\omega \in \Omega$; where ω is assumed to be outside the firm's control, and

Ω is the space of possible input market conditions.

The usual behavioural assumption made is that the firm chooses υ in such a way as to maximise profits; that is, given $(\overline{\rho},\overline{\omega}) \in P \times \Omega$, the firm chooses $\upsilon^* \in \Re^n_+$ so as to satisfy:

$$\text{for all } \upsilon \in \Re^n_+, R[\phi(\upsilon), \overline{\rho}] - K(\upsilon,\overline{\omega}) \underline{\leq} R[\phi(\upsilon^*), \overline{\rho}] - K(\upsilon^*,\overline{\omega}). \quad (1)$$

In what follows, we shall say that v* maximizes profits for ϕ, given $(\overline{\rho},\overline{\omega})$, iff υ* satisfies (1).

Define

$$\Re_{++} = \{x \in \Re | x > 0\}$$

and

$$\Re^n_{++} = \{\mathrm{w} \in \Re^n | \mathrm{w}_i > 0 \quad \text{for } i = 1, \ldots . n\}.$$

In this essay we shall assume that P and Ω are non-empty subsets of

$$P^* = \{\rho | \rho : \Re_+ \to \Re_{++}\} \text{ and } \Omega^* = \{\omega | \omega : \Re^n_+ \to \Re^n_{++}\},$$

respectively; and that R and K take the form

$$R(x, \rho) = \rho(x) \cdot x \quad \text{for } (x, \rho) \in \Re_+ \times P, \quad (2)$$

and

$$K(\upsilon, \omega) - \omega(\upsilon) \cdot \upsilon + C_0 \quad \text{for} (\upsilon, \omega) \in \Re^n_+ \times \Omega, \quad (3)$$

where $C_0 \geqq 0$ is the firm's fixed cost. Thus ρ and ω are, essentially, inverse demand (for output) and supply price functions for inputs, respectively.

The basic idea of this representation, as it applies to demand functions, is as follows. Under the usual assumptions of microeconomic theory, the demand function for x can be written as:

$$x = d(p_x, p, m, \alpha),$$

where

p_x = the price of good x,

$p = (p_1, \ldots p_l)$ is the vector of prices of other commodities in the economy,

$m = (m_1, \ldots m_k)$ is the vector of consumer incomes,

α is a parameter representing 'taste',

If, for each (p, m, α), $d(\cdot, p, m, \alpha)$ is strictly decreasing in p_x, we can invert the function to write

$$p_x = D(x, p, m, \alpha)$$

Each specification of (p, m, α) then determines a function $\rho : \Re_+ \to \Re_+$ defined by:

$$\rho(x) = D(x, p, m, \alpha) \quad \text{for } x \in \Re_+.$$

It is this sort of interpretation we have in mind by representing output market conditions by $\rho \in P^*$ and similar considerations apply to $\omega \in \Omega^*$. In accordance with this interpretation, we shall refer to elements of P* and Ω* as *output price* and *input supply price functions*, respectively.

Of particular importance to the analysis, however, is the case wherein the firm does not, by its own actions, change the market prices of the inputs which it uses; or the manager of the firm behaves as if this were the case (and thus is a 'price-taker' in input markets). In our formulation, this amounts to the assumption that

$$\Omega = \Omega^c, \quad (4)$$

where Ω^c is that subset of Ω* consisting of all constant functions (*constant supply price functions*) on $\Re^n_+$. In this situation $\omega \in \Omega^c$ corresponds to a unique value of $w \in \Re^n_{++}$ (and conversely);

and thus by a slight abuse of our notation, we can write the cost function (3) in the form:

$$K(\upsilon, w) = w \cdot \upsilon + C_0 \quad \text{for} (\upsilon, w) \in \Re^n_+ \times \Re^n_{++}. \tag{5}$$

Since our principal concern will be with the theory of pure competition, we shall assume throughout our discussion of the behaviour of the individual firm that (4) and (5) hold. However, the more general specification of K will be useful when we turn to the discussion of competitive equilibrium for an industry.

When (4) and (5) hold, it is useful to break the firm's profit-maximization problem down into two parts, as follows.

1. Defining $X \subseteq \Re_+$, the firm's *producible set*, by:

$$X = \left\{x \in \Re_+ | \left(\exists \upsilon \in \Re^n_+\right) : \phi(\upsilon) \geqq x\right\}, \tag{6}$$

we find, for each $(\mathrm{x}, \mathrm{w}) \in X \times \Re^n_{++}$, $c(\mathrm{x}, \mathrm{w})$, the minimum (variable) cost of producing x, given w. That is, given $(\mathrm{x}^*, \mathrm{w}^*) \in X \times \Re^n_{++}$, we find $\upsilon^* \in \Re^n_+$ satisfying:

$$\phi(\upsilon^*) \geqq x^* \text{ and } w^* \cdot \upsilon^* \leqq w^* \cdot \upsilon \text{ for all } \upsilon \in \Re^n_+ \text{ such that } \phi(\upsilon) \geqq x^* \tag{7}$$

Variable cost and total cost at (x*, w*), $c(x^*, w^*)$ and $c(x^*, w^*)$, respectively, are then given by:

$$c(x^*, w^*) = w^* \cdot v^*,$$

and

$$C(x^*, w^*) = c(x^*, w^*) + C_0$$

2. Given the function C, and $\rho \in P$, *we* then find $x^* \in X$ satisfying

$$\text{for all } x \in X, R(x, \rho) - C(x, w) \leqq R(x^*, \rho) - C(x^*, w^*); \tag{8}$$

or equivalently,

$$\text{for all } x \in X, R(x, \rho) - c(x, w) \leqq R(x^*, \rho) - c(x^*, w^*); \tag{8'}$$

Because of the equivalance of (8) and (8'), we shall hereafter concern ourselves only with the variable cost function, c, and we shall refer to it as simply the 'cost function'. Similarly we define the firm's (gross) *profit function*, $\pi : X \times P \times \Re^n_{++} \to R$ by

$$\pi(x, \rho, w) = R(x, \rho) - c(x, w); \tag{9}$$

and we shall say that $x^* \in X$ maximizes the profit function $\pi(\cdot, \rho, w)$ if, and only if, it satisfies(8′)

It can be shown that if $R(\cdot, \rho)$ is non-decreasing in x on X, and υ* maximizes profits for ϕ, given (ρ, w), then

$$w \cdot \upsilon^* = c[\phi(\upsilon^*), w], \tag{10}$$

and $x^* \equiv (\upsilon^*)$ satisfies (8′). Conversely, if $x^* \in X$ satisfies (8′), and υ* satisfies (7), then υ* maximizes profits for ϕ, given (ρ, w). Thus, if $R(\cdot, \rho)$ is non-decreasing on X; the two procedures are logically equivalent. If $R(\cdot, \rho)$ is decreasing on a portion of X (the case of inelastic demand), the two-step procedure is not necessarily equivalent to maximizing profits for ϕ, given (ρ, w). However, in the situation we will be examining in detail, $R(\cdot, \rho)$ will always be non-decreasing on X.

One advantage of the two-step analysis of profit maximization is that the problem of maximizing the profit function, π, is much simpler, both conceptually and operationally, than the problem of finding $\upsilon^* \in \Re^n_+$ which maximizes profits for ϕ. However, a much more significant advantage of the two-step analysis is that the analysis of the cost function itself is applicable whatever the form of the revenue function, R. In particular, the relationship between a firm's production and cost functions, the examination of which will be our next order of business, is of interest whatever the (output) market structure under investigation.

In order to formally analyse the relationship between production and cost, let us say that a function $\phi : \Re^n_+ \to \Re_+$ is a *production function* if, and only if, $\phi(\cdot)$ satisfies the following three properties:

P.1 $\phi(0) = 0$ and for some, $\upsilon^{T} \in \Re^{n}_{+}$, $\phi(\upsilon^{T}) > 0$.
P.2 ϕ is non-decreasing on $\Re^{n}_{+}$, i.e.,
for each $\upsilon, \upsilon' \in \Re^{n}_{+}$, if $\upsilon \geqq \upsilon'$, then $\phi(\upsilon) \geqq \phi(\upsilon')$.
P.3 ϕ is upper semi-continuous on $\Re^{n}_{+}$.

In the following treatment, we shall also define:

$$X_{+} = \{x \in X | x > 0\} = \left\{x \in \Re_{+} | \left(\exists \upsilon \in \Re^{n}_{+}\right) : 0 < x \leqq \phi(\upsilon)\right\}$$

and

$$\Re^{n}_{++} = \left\{w \in \Re^{n}_{+} | w_i > 0 \quad \text{for } i = 1, ..., n\right\}.$$

The above conditions are generally fairly standard, and probably require no discussion, except perhaps the assumption that ϕ is upper semi-continuous. Generally speaking, continuity is *not* an empirically meaningful condition; that is, in investigating an actual production process we can do no better than to observe a finite collection of input vectors $\{v^1, \ldots, v^\tau\}$, together with the associated values of output $\{x^1, \ldots, x^\tau\}$. If these observations are consistent with the hypothesis that there exists a function $\phi : \Re^{n}_{+} \to \Re_{+}$ such that

$$x^t = \phi(x^t) \quad \text{for } t = 1, \ldots, \tau; \tag{11}$$

then they are also consistent with the additional assumption that ϕ is continuous. Consequently, one might very legitimately ask why I have not simply made the more familiar assumption that ϕ is continuous. My reason for not doing so is that there are situations in which we may want to interpret the quantity of output as a stock, rather than a flow. In some situations where this is the case (e.g., 'finite production runs'), it is likely to be appropriate to assume that ϕ is discrete-valued; and while it is possible for a function $\phi : \Re^{n}_{+} \to \Re_{+}$ to be both upper semi-continuous and discrete-valued (an example is presented later in the text), it is not possible for such a function to be both discrete-valued and continuous. As it turns out, most of the analysis to follow requires only that ϕ be upper semi-continuous, and thus is applicable with either a stock or a flow interpretation of output quantities.

Turning now to the derivation of the cost function, let $(x^*, w^*) \in X \times \Re^{n}_{++}$. Since ϕ is upper semi-continuous, it can be shown that there exists $\upsilon^* \in \Re^{n}_{+}$ such that υ^* satisfies (7) for $x = x^*$ and $w = w^*$. Thus it follows that if ϕ is a production function, by our definition, then the (variable) cost function $c : X \times \Re^{n}_{++} \to \Re_{+}$ is well-defined. In fact, it is fairly easy to show that the cost function $c(\cdot)$ corresponding to a given production function, $\phi(\cdot)$, will satisfy the following conditions. (Most of these properties are established in McFadden, 1978, and those which are not are proved in Moore, 1986.)

C.1. X, the producible set for ϕ, is a sub-interval of $\Re_{+}$, and either
  a. there exists $\overline{x} > 0$ such that $X = [0, \overline{x}]$, or
  b. for each $w \in \Re^{n}_{++}$, $c(\cdot, w)$ is unbounded on X.
C.2. for each $w \in \Re^{n}_{++}$:
  a. $c(\cdot, w)$ is non-decreasing in x,
  b. $c(0, w) = 0$,
  c. $c(x, w) > 0$ for each $x \in X_{+}$,
  d. $c(\cdot, w)$ is lower semi-continuous on X.
C.3. for each $x \in X_{+}$, $c(x, \cdot)$ is:
  a. increasing in w.
  b. positively homogeneous of degree one in w.
  c. concave in w.
  d. continuous in w.

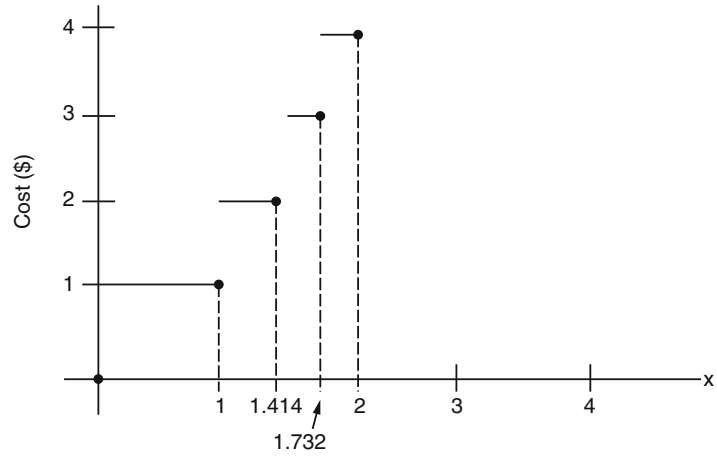
Since $C(\cdot)$, the *total cost function for* ϕ is defined by

$$C(w) = c(w) + K_0 \quad \text{for } (w) \in \Re^{n}_{++}, \tag{12}$$

it follows at once that C also satisfies all of the above properties *except* C.2.b and C.3.b. We can also define the *average variable cost*, and *average cost* functions on $X_{+} \times \Re^{n}_{++}$ by:

$$a(w) = c(w)/x \quad \text{for}(w) \in X_{+} \times \Re^{n}_{++}, \tag{13}$$

Cost and Supply Curves, Fig. 1



and

$$A(\mathrm{x},\mathrm{w}) = C(\mathrm{x},\mathrm{w})/w \quad \text{for}(\mathrm{x},\mathrm{w}) \in X_+ \times \Re^{n}_{++}, \tag{14}$$

respectively. However, under the assumptions we have been employing to this point, the marginal cost function, $c_x(\cdot, w)$, will not necessarily be well-defined. ['$c_x(x, w)$' denotes the partial derivative of c with respect to x, evaluated at (x, w)]. In fact, under the assumptions which we have been employing, a cost function may look rather unlike those used in traditional intermediate theory diagrams, as will be seen from the following example.

For each $x \in \Re$, define $\lfloor x \rfloor$ and $\lceil x \rceil$ by:

$\lfloor x \rfloor$ = that unique integer, n, satisfying $\mathrm{n} \leqq \mathrm{x} < \mathrm{n} < +1$,

and

$\lceil x \rceil$ = that unique integer, n', satisfying:

$$\mathrm{n}' - 1 < \mathrm{x} \leqq \mathrm{n}'.$$

If we let $\alpha > 0$ be a positive real number, and define $\phi : \Re_+ \to \Re_+$ by

$$\phi(x) = \alpha\sqrt{\lfloor \upsilon \rfloor}, \quad \text{for } \upsilon \in \Re_+,$$

it can be shown that ϕ is a production function. Moreover, it is easy to see that the producible set for ϕ is given by:

$$X = \Re_+,$$

and that the cost function for ϕ is given by:

$$c(x, w) = w\left\lceil (x/\alpha)^2 \right\rceil \quad \text{for } x \in X \quad \text{and} \quad w \in \Re_{++}.$$

The graph of $c(\cdot)$ for $\alpha = w = 1$ is shown in Fig. 1, below.

Obviously, the marginal cost function is not defined in this example.

Suppose now that the firm sells its output in a competitive market; that is, that the firm is a 'price-taker' in its output market. In terms of our formulation, this means that P is the set of all constant functions (*constant output price functions*); and by considerations similar to those used in our development of the cost function, we can write the firm's revenue function as:

$$\mathrm{R}(\mathrm{x}, p) = \mathrm{p} \cdot \mathrm{x} \quad \text{for } (\mathrm{x},\mathrm{p}) \in X \times \Re_{++}. \tag{15}$$

Thus we are interested in whether or when, given $(p, w) \in \Re_{++} \times \Re^{n}_{++}$, there exists $x^* \in X$ satisfying

$$\begin{aligned} &\text{for all } x \in X, \\ &p \cdot x - c(x, w) \leqq p \cdot x^* - c(x^*, w). \end{aligned} \tag{16}$$

If, for each $(p, w) \in \Re_{++} \times \Re^{n}_{++}$, there exists a unique $x^* \in X$ satisfying (16), then there exists a function $s : \Re_+ \times \Re^{n}_{++} \to X$ such that for each

(p, w), $s(p, w)$ is that unique value of x, x^*, satisfying (16). This function is the firm's *supply function*; its value at each (p, w) is the firm's competitive (profit-maximizing) output, given the output, price, p, and the vector of input prices, w. We shall briefly examine the conditions under which this function will exist, and its relationship to the firm's cost curve.

Suppose for the moment that the cost function is differentiable in x, and let '$c_x(x, w)$' denote the partial derivative of c with respect to x, evaluated at $(\mathrm{x}, \mathrm{w}) \in X \times \Re^n_{++}$. The function $c_x(\cdot)$ is called the firm's *marginal cost function*, and it is customary in intermediate microtheory texts to state that the firm's supply curve is 'that portion of the marginal cost curve lying above the average variable cost curve'. Clearly the situation is a bit more complicated than this; and these complications, while perhaps not critical in the analysis of the individual competitive firm, become troublesome when we turn to the analysis of supply in a competitive industry.

In examining this problem, we shall restrict our attention to the case in which the text book treatment mentioned above is most nearly correct (or at least the most favorable case I can come up with); namely, that in which the firm's cost function satisfies the following condition.

C.4. Defining $\chi = \sup X$ (χ may, of course, be equal to $+\infty$), c is continuously differentiable in x on $[0, \chi[\times \Re^n_{++}$; and, for each $w \in \Re^n_{++}$, $c_x(\cdot)$ is strictly increasing on $[0, \chi[$. [By '$c_x(0, w)$' we mean the right hand partial derivative.

We shall not attempt to develop conditions on the firm's production function sufficient to ensure that c satisfies C.4. However, it is easy to show that if ϕ is strictly concave, then $c(\cdot, w)$ will be strictly convex in x, for each $w \in \Re^n_{++}$. If in this case c is differentiable as well, then $c_x(\cdot, w)$ will be strictly increasing in x on $[0, \chi[$, for each $w \in \Re^n_{++}$. Moreover, it follows from recent results on duality relationships between production and cost-functions that, given any cost function, c, satisfying (C.1–C.3 and) C.4, there exists a unique concave production function for which c is the corresponding cost function. (For an excellent survey of duality, see Diewert, 1982.)

Suppose, then, that c satisfies C.4. It is easy to see that, for each $\mathrm{w} \in \Re^n_{++}$,

$$\lim_{x \to \chi} c_x(x, w)$$

exists, although it may be equal to $+\infty$. Furthermore, if for each $w \in \Re^n_{++}$, we define $a(w)$ and $b(w)$ by

$$a(w) = c_x(0, w) \quad b(w) = \lim_{x \to \chi} c_x(x, w);$$

it follows from C.4 that for each $p \in [a(w), b(w)]$, there exists a unique $x \in [0, \chi]$ such that

$$p = c_x(x, w).$$

Consequently, $c_x(\cdot, w)$ is invertible (in x) on the set π^0 defined by

$$\pi^0 = \{(\mathrm{p}, \mathrm{w}) \in \Re_{++} \times \Re^n_{++} \mid p \in [a(\mathrm{w}), b(\mathrm{w})]\};$$

that is, there exists a function $\sigma : \pi^0 \to [0, \chi] \subseteq X$ satisfying:

$$\text{for each } (p, w) \in \pi^0, \quad c_x[\sigma(p, w), w] = p, \tag{17}$$

and

$$\text{for each } (\mathrm{x}, \mathrm{w}) \in [0, \chi[\times \Re^n_{++}, \ \sigma[c_x(\mathrm{x}, \mathrm{w}), \mathrm{w}] = x \tag{18}$$

If we define π as the set of all $(\mathrm{p}, \mathrm{w}) \in \Re_{++} \times \Re_{++}$ such that there exists $x^* \in X$ satisfying (16), the interested reader should have no great difficulty in proving the following.

S.1. $\pi^0 \subseteq \pi$

S.2. For each $(p, w) \in \pi$, the profit-maximizing output is unique; thus the firm's supply function has domain π, and is given by:

$$s(p, w) = \left\{ \begin{array}{ll} 0 & \text{if } 0 < \mathrm{p} \leqq a(\mathrm{w}) \\ \sigma(\mathrm{p}, \mathrm{w}) & \text{if } p \in]\mathrm{a}(\mathrm{w}), \mathrm{b}(\mathrm{w})[\\ \chi & \text{if } p \geqq b(w) \end{array} \right\} \text{for} (\mathrm{p}, \mathrm{w}) \in \pi \tag{19}$$

S.3. For each $w \in \mathfrak{R}^n_{++}$, $s(\cdot, w)$, is continuous and non-decreasing in p; and is strictly increasing in p on $]a(w), b(w)[$.

While, as already indicated, I shall leave the proof of the above result to the 'interested reader'; one or two explanatory comments seem to be in order. First, if c satisfies C.4, then marginal cost is always at least as large as average variable cost, i.e.,

$$\text{for each } (\mathrm{x}, \mathrm{w}) \in X_+ \times \mathfrak{R}^n_{++}, \quad c_x(\mathrm{x}, \mathrm{w}) > c(\mathrm{x}, \mathrm{w})/x \tag{20}$$

(this is easily established by the use of the mean value theorem). Secondly, the condition $p \geqq b(w)$ and $(p, w) \in \pi$ is a bit misleading. The fact of the matter is, these two conditions can hold simultaneously only if X is of the form $X = [0, \chi]$, with χ finite (and positive. In this case also $\pi = \mathfrak{R}_{++} \times \mathfrak{R}^n_{++}$.) If this is not the case, then by C.1 we see that two other cases are possible.

Case 1: $X =]0, \chi[$ with χ finite. Here we have by C.1 that $c(\cdot, w)$ is unbounded in X, and this fact can be used to show that $b(w) = +\infty$, for each $w \in \mathfrak{R}^n_{++}$. Thus in this case, $\pi = \mathfrak{R}_{++} \times \mathfrak{R}^n_{++}$, but we cannot have $p \geq b(w)$.

Case 2: $X = \mathfrak{R}_+$. Here it is possible that $b(w)$ is finite. However, if this is the case, and (p, w) is such that $p \geqq b(w)$, then no profit-maximizing solution exists. For example, consider the homothetic cost function

$$c(x, w) = f(x)\gamma(w), \tag{21}$$

where γ can be any function satisfying C.3, and f is given by:

$$f(x) = \alpha x + \left(x^2 + \beta^2\right)^{1/2} - \beta, \quad \text{with } \alpha, \beta > 0. \tag{22}$$

Here it is easily shown that c satisfies C.1–C.4, and that

$$a(w) = c_x(0, w) = f'(x)\gamma(w) = a\gamma(w), \tag{23}$$

While

$$\begin{aligned} b(w) &= \lim_{\chi \to +\infty} c_x(x, w) = \gamma(w) \lim_{x \to +\infty} f'(x) \\ &= (\alpha + 1)\gamma(w). \end{aligned} \tag{24}$$

Here $b(w)$ is finite, but if $p \geqq b(w)$, then no profit-maximizing output exists.

Turning our attention to suppy conditions for a competitive industry, suppose there are m firms producing a single (homogeneous) commodity, and let the ith firm's production function, cost function, and producible set be denoted by ϕ^i, c^i, and X_i, respectively. We shall use the generic notation x_i and v^i to denote the ith firm's output and vector of inputs employed, respectively; and we define the *producible set for the industry*, X, by

$$X = \sum_{i=1}^{m} X_i.$$

We assume that the market is competitive, so that each firm is a price-taker in both output and input markets; and we shall be interested in the competitive equilibria of the market (or industry), defined as follows.

We shall say that $\left(\overline{v}^i, \ldots, \overline{v}^m; \overline{\rho}, \overline{\omega}\right)$ is a *competitive equilibrium for the industry* iff.

1. $\overline{v}^i \in \mathfrak{R}^n_+$ for $i = 1, \ldots, m$,
2. $\overline{\rho} \in P^*, \overline{\omega} \in \Omega^*$, and
3. defining $\overline{p} = \overline{\rho}\left[\sum_{i=1}^{m} \phi^i\left(\overline{v}^i\right)\right]$ and $\overline{w} = \overline{\omega}\left(\sum_{i=1}^{m} \overline{v}^i\right)$,

the following condition holds: for each $i (i = 1, \ldots, m)$, $\overline{v}^i$ satisfies

$$\text{for all } v^i \in \mathfrak{R}^n_+, \overline{p} \cdot \phi^i(v^i) - \overline{w} \cdot v^i \underline{\leq} \overline{p} \cdot \phi^i\left(\overline{v}^i\right) - \overline{w} \cdot \overline{v}^i \tag{25}$$

We shall also be interested in aggregate competitive equilibria for the industry, defined as follows.

We shall say that $(\overline{x},\overline{v},\overline{p},\overline{w}) \in X \times \mathfrak{R}^n_+ \times P^* \times \Omega^*$ is an *aggregate competitive equilibrium for the industry* iff there exist $\overline{v}^i \in \mathfrak{R}^n_+$ for $i = i, \ldots, m$ such that

$$\overline{v} = \sum_{i=1}^{m} \overline{v}^i, \quad \overline{x} = \sum_{i=1}^{m} \phi^i(\overline{v}^i),$$

and $(\overline{v}^1, \ldots, \overline{v}^m; \overline{p}, \overline{\omega})$ is a competitive equilibrium for the industry. We shall say that $(\overline{x},\overline{p},\overline{\omega}) \in X \times P^* \times \Omega^*$ is an aggregate output equilibrium for the industry iff there exists $\overline{v} \in \mathfrak{R}^n_+$ such that $(\overline{x},\overline{v},\overline{p},\overline{\omega})$ is an aggregate competitive equilibrium for the industry.

The first question I would like to consider is whether we can characterize the aggregate output equilibria as the set of points where supply equals demand, if by 'supply' we mean the summation of the supply functions of the individual firms. I don't believe that it will be necessary to go into a lot of detail to convince the reader of the truth of the following two assertions.

1. We cannot obtain the set of output equilibria for the industry by summing the individual supply curves and then finding the set of points at which this summation function equals demand unless we restrict our attention to cases in which $\omega \in \Omega^c$. This is a fairly serious difficulty, since this assumes that all m firms in the industry can simultaneously expand or contract input usage without appreciably affecting input prices, an assumption which is much more restrictive than assuming that each firm behaves as if its individual actions have no appreciable effect on input prices.
2. Even if we restrict our attention to the case where $\Omega = \Omega^c$, the analysis of the form of the summation supply function will be a very messy business. Suppose, for example, that each cost function c^i satisfies C.4, and denote the ith firm's supply function by s^i and its domain by π^i. Then the summation (industry) supply function will be defined on

$$\pi = \bigcap_{i=1}^{m} \pi_i,$$

and will be given by

$$s(p,w) = \sum_{i=1}^{m} s^i(p,w) \quad \text{for } (p,w) \in \pi.$$

Given $\overline{w} \in \mathfrak{R}^n_{++}$ and $\rho \in P^*$, $(\overline{x},\overline{p},\overline{\omega})$ will be an aggregate output equilibrium for the industry in this case if, and only if,

$$[\overline{p}(\overline{x}), \overline{w}] \in \pi \quad \text{and} \quad x = s[\overline{p}(\overline{x}), \overline{w}]. \tag{26}$$

In this case, therefore one can obtain the industry supply function by summing the supply functions of the individual firms, and this industry supply function can be conveniently utilized to obtain the aggregate output equilibria for the industry. However, if the reader will re-examine condition S.2, above [and particularly equation (19)], I am sure you will have no difficulty convincing yourself that the characterization of the form of π and $s(\cdot)$ will be a very messy business; even under the somewhat restrictive assumption that each c^i satisfies C.4. In fact, this is a fairly substantive problem, for suppose one wishes to estimate the supply function for a purely competitive industry. Given our present estimation techniques, this means that one needs to specify, a functional form for $s(\cdot)$ (actually, a parametric family, the individual elements of which are determined by a finite set of parameters), whose parameters can be estimated from data on aggregate output equilibria for the market. The question is, what sort of functional form is appropriate? Once again, a glance back at condition S.2 and equation (19) will suffice to convince the reader that the only practical means of deriving the form $s(\cdot)$ from assumptions about the functional form of the individual cost functions, c^i, is to assume that all of the individual firms have the same cost function; an assumption with which it is difficult to be comfortable.

Fortunately, there is a way of circumventing both of these difficulties, at least to a great extent. The relationships we shall develop to facilitate this alternative analysis are in some sense 'well known', but for some reason they do not seem to have found their way into microeconomic theory textbooks. We begin by considering the following problem, for an arbitrary $v \in \mathfrak{R}^n_+$:

C

$$\text{Maximize} \sum_{i=1}^{m} \phi^i(v^i), \text{subject to} \sum_{i=1}^{m} v^i \leqq v \text{ and } v^i \in \Re_+^n \text{ for } i = 1, .., m. \tag{27}$$

Since each ϕ^i is upper semi-continuous, the function

$$F(v^1, \ldots, v^m) \equiv \sum_{i=1}^{m} \phi^i(v^i)$$

is also upper semi-continuous; and, for each $v \in \Re_+^n$, the set

$$V(v) = \left\{ (v^1, \ldots, v^m) \in \Re^{mn} \middle| \sum_{i=1}^{m} v^i \leqq v \right\}$$

is compact. Therefore, for each $v \in \Re_+^n$, the problem (27) has a solution, the value of which we shall denote by '$\phi(v)$', that is,

$$\phi(v) = \max \left\{ \sum_{i=1}^{m} \phi^i(v^i) \mid (v^1, \ldots, v^m) \in V(v) \right\}. \tag{28}$$

We shall refer to the function ϕ as the *Industry production function*; and for each $v \in \Re_+^n$ we shall denote the solution set of (27) by '$V^*(v)$', that is,

$$V^*(v) = \left\{ (v^1, \ldots, v^m) \in V(v) \middle| \sum_{i=1}^{m} \phi^i(v^i) = \phi(v) \right\}. \tag{29}$$

It can be shown that ϕ is a production function; that is, it satisfies P.1–P.3. The industry production function does not necessarily retain other properties of the individual production functions, ϕ^i; for example, it may be the case that all the ϕ^i are quasi-concave but that ϕ is not quasi-concave. On the other hand, if all the ϕ^i are concave, or are all positively homogeneous of degree $r > 0$, then ϕ is a production function, and thus has a cost function satisfying C.1–C.3. We can show that the set of aggregate competitive equilibria for the industry are the solutions of the following problem

Given a production function, ϕ, and $(\overline{p}, \overline{\omega}) \in P^* \times \Omega^*$, we shall say that $\overline{v} \in \Re_+^n$ *myopically maximizes profits for* ϕ, *given* $(\overline{p}, \overline{\omega})$ if $\overline{v}$ satisfies:

$$\text{for all } v \in \Re_+^n, \overline{p} \cdot \phi(v) - \overline{w} \cdot v \leqq \overline{p} \cdot \phi(\overline{v}) - \overline{w} \cdot \overline{v}, \tag{30}$$

where $\overline{p} = \overline{p}[\phi(\overline{v})]$ and $\overline{w} = \overline{\omega}(\overline{v})$.

One can then prove the following result; although, in the interest of brevity, I shall not do so here (a proof is included in Moore, 1986).

Proposition 1 If $\overline{v} \in \Re_+^n$ myopically maximizes profits for ϕ (the industry production function), given $(\overline{p}, \overline{\omega}) \in P^* \times \Omega^*$, then $[\phi(\overline{v}), \overline{v}, \overline{p}, \overline{\omega}]$ is an aggregate competitive equilibrium for the industry [in fact, in this situation $(\overline{v}^1, \ldots, \overline{v}^m; \overline{p}, \overline{\omega})$ is a competitive equilibrium for the industry, for any $(\overline{v}^1, \ldots, \overline{v}^m) \in V^*(\overline{v})$]. Conversely, if $(\widehat{v}^1, \ldots, \widehat{v}^m; \widehat{\rho}, \widehat{\omega})$ is a competitive equilibrium for the industry, and we define

$$\widehat{v} = \sum_{i=1}^{m}, \widehat{v}^i,$$

then

$$\phi(\widehat{v}) = \sum_{i=1}^{m} \phi^i(\widehat{v}^i),$$

and $\widehat{v}$ myopically maximizes profits for ϕ, given $(\widehat{\rho}, \widehat{\omega})$.

There are a number of points which appear to be worth making with regard to the implications of the above result.

1. Given $(\overline{p}, \overline{\omega}) \in P^* \times \Omega^*$, one can find all aggregate competitive output and input equilibria corresponding to $(\overline{p}, \overline{\omega})$ by finding the set of v which myopically maximize ϕ, given $(\overline{p}, \overline{\omega})$. If there is a unique maximizing value of v, then there is correspondingly a unique aggregate competitive equilibrium for the industry. (This uniqueness question is explored in Moore, 1986).
2. One can estimate the industry production function, ϕ, from observations on aggregate output

and input usage associated with competitive equilibria for the industry; since if $(\overline{v}^1, \ldots, \overline{v}^m; \overline{p}, \overline{\omega})$ is a competitive equilibrium, for the industry, then

$$\sum_{i=1}^{m} \phi^i(\overline{v}^i) = \phi\left(\sum_{i=1}^{m} \overline{v}^i\right).$$

Similarly if one also has data on the values of w associated with each such competitive equilibrium, one can also estimate $\overline{\omega}$.

There is a complication connected with this last point, however. The situation I have been describing is one in which we have a series of competitive equilibria for the industry;

$$(x^t, v^t, \rho^t, \omega^t) \quad \text{for } t = 1, \ldots, T; \tag{31}$$

and we have observed

$$(x^t, v^t, w^t) \quad \text{for } t = 1, \ldots, T, \tag{32}$$

where

$$w^t = \omega^t(v^t) \quad \text{for } t = 1, \ldots, T. \tag{33}$$

Under the assumption that the individual production functions, ϕ^i, are unchanged over the period, then we can clearly estimate the industry production function from the relation

$$x^t = \phi(v^t) \quad \text{for } t = 1, \ldots, T, \tag{34}$$

where (34) holds by virtue of the fact that

$$x^t = \sum_{i=1}^{m} \phi^i(v^{it}), \quad v^t = \sum_{i=1}^{m} v^{it},$$

and

$$\sum_{i=1}^{m} \phi^i(v^{it}) = \phi\left(\sum_{i=1}^{m} v^{it}\right).$$

Assuming that there exists $\overline{w} \in \Omega^*$ such that

$$\omega^t = \overline{\omega} \quad \text{for } t = 1, \ldots T; \tag{35}$$

we can equally well estimate $\overline{w}$ from the relationship

$$w^t = \omega^t(v^t) = \overline{\omega}(v^t) \quad \text{for } t = 1, \ldots, T. \tag{36}$$

However, here it is well to keep in mind a point raised by Joan Robinson in her famous article 'Rising Supply Price' (1941): if the industry output price function changes from, say, ρ^1 to ρ^2, there will generally be a corresponding change in the input supply price function as well (in our framework, a change from some ω^1 to $w^2 \in \Omega^*$). In the situation under consideration here, and under the assumption that the ϕ^i's are unchanged, we must have $\rho^t \neq \rho^{t'}$ if $(x^t, v^t) \neq (x^t, v^t)$. Thus we have to allow for the possibility that $\omega^t \neq w^{r'}$ as well (unless, of course, $\omega^t = \omega^{r'}$ for all $t, t' = 1, \ldots, T$). This is essentially, an identification problem, and can be handled by methods similar to those used in conventional demand and supply estimation [on this, see, e.g., Fisher (1966)].

Robinson's point is also pertinent to the predictive use of the methods we have discussed here, however. Suppose we know, or have estimated, both the industry production function, ϕ, and the prevailing input supply price function $\overline{\omega}$; and we wish to predict the aggregate competitive equilibrium associated with $\widehat{\rho} \in P^*$. It follows from Proposition *1* that we can find $(\overline{x}, \overline{v}, \widehat{\rho}, \overline{\omega})$ by finding $\overline{v} \in \Re^n_+$ such that $\overline{v}$ myopically maximizes profits for ϕ, given $(\widehat{\rho}, \overline{\omega})$. The problem here, of course, is that if demand conditions change to $\widehat{\rho}$, there may be a corresponding change to a new input supply price function, $\widehat{\omega}$. While I have no suggestions regarding a way of circumventing this difficulty, one suspects that there are many more situations in which it is safe to assume that industry supply conditions remain unchanged for 'small' changes in ρ, than there are in which input prices themselves remain fixed for such changes in ρ.

3. Let us call the cost function, c, corresponding to the industry production function, ϕ, the *industry cost function*. It can be shown that if

$(\overline{v}^1, \ldots, \overline{v}^m; \overline{p}, \overline{\omega})$ is a competitive equilibrium for the industry, and we define

$$\overline{w} = \overline{\omega}\left(\sum_{i=1}^{m} \overline{v}^i\right), \quad \text{and} \quad \overline{x}_i = \phi^i\left(\overline{v}^i\right) \text{for } i = 1, \ldots, m,$$

then

$$\sum_{i=1}^{m} \overline{w} \cdot v^i = c(\overline{x}, \overline{w}), \quad \text{where} \quad \overline{x} = \sum_{i=1}^{m} \overline{x}_i.$$

Thus one can estimate the industry cost function by observations on aggregate industry output, total industry factor cost, and $\overline{w}$. This fact is of particular interest in connection with our next point.

4. In the constant input supply price case $(\Omega = \Omega^c)$, our analysis yields a simplified approach to obtaining (and analyzing the properties of) the industry supply function. From Proposition *1* and our discussion of profit maximization by the individual firm, we can see that if we derive the function s: $\pi \rightarrow X$ from the industry cost function from its cost function (see the derivation of S.1–S.3 above), we can find the aggregate output equilibrium for the industry, given $(\overline{p}, \overline{w})$, by finding $\overline{x} \in X$ satisfying

$$\overline{x} = s\left[\overline{p}(\overline{x}), \overline{w}\right).$$

In other words, it is legitimate to approach the problem of estimating a supply curve for an industry by specifying a production function (or cost function) for the industry were a as an aggregate, and proceeding as if the industry were a single profit-maximizing entity.

The perceptive reader may, however, have detected a problem with the above procedure. We have shown that if each firm in a competitive industry has an individual production function (by the definition used here), then there is an associated total product function for the industry which is also a production function. The question remains, however, whether *any* production function could serve to describe industry output relationships in this manner. Insofar as I am aware, the answer to this general question is unknown, and deserves investigation. However, any *concave* production function can play such a role, as is shown by the following.

Proposition 2 Let: $\phi : \Re_+^n \rightarrow \Re_+$ be a concave production function, and let m be a positive integer. Then there exist concave production functions: $\phi^i : \Re_+^n \rightarrow \Re_+$ for $i = 1, \ldots, m$, such that ϕ is the industry production function associated with the ϕ^i.

While I shall not provide a complete proof of Proposition *2* here, it can be established by showing that if ϕ is a concave production function, and one defines $\phi = (i = 1, \ldots, m)$ by

$$\phi^i(v^i) = (1/m)\phi(mv^i) \quad \text{for } v^i \in \Re_+^n;$$

then ϕ^i is a concave production function, and ϕ is the industry production function associated with $(\phi^i, \ldots, \phi^m)$. (Cf. Stigum 1986).

See also

► Cost Functions

References

Diewert, W.E. 1982. Duality approaches to microeconomic theory. In *Handbook of mathematical economics*, vol. II, ed. K.J. Arrow and M.D. Intriligator, 535–599. Amsterdam: North-Holland.

Fisher, F.M. 1966. *The identification problem in econometrics*. New York: McGraw-Hill.

Jacobsen, S.E. 1970. Production correspondences. *Econometrica* 38(5): 754–770.

Jacobsen, S.E. 1972. On Shephard's duality theorem. *Journal of Economic Theory* 4(3): 458–464.

McFadden, D. 1978. Cost, revenue, and profit functions. In *Production economics: A dual approach to theory and applications*, ed. M. Fuss and D. McFadden, 3–109. Amsterdam: North-Holland.

Moore, J.C. 1986. *A reconsideration of market supply and demand analysis*. Purdue University, Mimeo.

Robinson, J. 1941. Rising supply price. *Economic*, N.S. 8: 1–8.

Shephard, R.W. 1970. *Theory of cost and production functions*. Princeton: Princeton University Press.
Stigum, B.P. 1986. On a property of concave functions. *Review of Economic Studies* 35(4): 413–416.
Varian, H.R. *Microeconomic analysis*, 2nd ed. New York: Norton.

Cost Functions

W. Erwin Diewert
Organization Name, City, UK

Keywords
Cost functions; Duality; Euler's theorem; Expenditure functions; Generalized Leontief cost function; Indirect utility function; Normalized quadratic cost function; Production functions; Shephard's duality theorem; Shephard's lemma; Substitutes and complements; Translog cost function; Unit cost function

JEL Classifications
D2

Introduction

Cost and expenditure functions are widely used in both theoretical and applied economics. Cost functions are often used in econometric studies which describe the technology of firms or industries while their consumer theory counterparts, expenditure functions, are frequently used to describe the preferences of consumers.

Cost and expenditure functions also play an important role in many theoretical investigations. This is due to the fact that a cost function embodies the consequences of cost minimizing behaviour on the part of a consumer or producer and so it is not necessary to spell out the details of the primal minimization problem that defined the cost function. This may seem like a very minor advantage, but when one is dealing with, say, the comparative statics of a general equilibrium problem, the use of cost functions leads to the analysis of a much smaller system of equations and hence the structure of the problem can be more easily understood.

Sections "Properties of Cost Functions," "Duality Between Cost and Production Functions," "The Derivative Property of the Cost Function," and "The Comparative Statics Properties of Cost Functions" below develop the theoretical properties of cost functions while sections "Functional Forms for Cost Functions," "Applications to the Estimation of Consumer Preferences," and "Cost Functions and Measures of Welfare Gain" are devoted to empirical applications of cost functions in the producer and consumer contexts.

Properties of Cost Functions

One of the fundamental paradigms in economics is the one which has a producer competitively minimizing costs subject to his technological constraints. Competitive means that the producer takes input prices as fixed during the given period of time irrespective of the producer's demand for those inputs.

We assume that only one output can be produced using N inputs and that the producer's technology can be summarized by a *production function* $F : y = F(x)$ where $y \geq 0$ is the maximal amount of output that can be produced during a period, given the non-negative vector of inputs $x \equiv (x_1, \ldots, x_N) \geq 0_N$. We further assume that the cost of purchasing one unit of input i is $p_i > 0$, $i = 1, \ldots, N$ and that the positive vector of input prices that the producer faces is $p \equiv (p_1, \ldots, p_N) \gg 0_N$.

For $y \geq 0$, $p \gg 0_N$, the producer's *cost function* C is defined as the solution to the following constrained minimization problem:

$$C(y,p) \equiv \min_x \{p.x : F(x) \geq y, x \geq 0_N\} \qquad (1)$$

where $p \cdot x \equiv \sum_{n=1}^{N} p_n x_n$. Thus $C(y, p)$ is the minimum input cost of producing at least the output level y, given that the producer faces the input price vector p.

The minimization problem (1) can also be given a consumer theory interpretation: let F be a consumer's preference or *utility function,* let y be a utility or welfare level, let x be a vector of commodity purchases (rentals in the case of consumer durables), and let p be a vector of commodity (rental) prices. In this case, the consumer attempts to minimize the cost of achieving at least the target welfare level indexed by y, and the solution to (1) defines the consumer's *expenditure function.*

Unfortunately, the minimum (1) may not exist in general. However, if we impose the following very weak regularity condition on F, it can be shown that C will be well defined as a minimum: *Assumption* 1 *on F: F* is continuous from above.

Assumption 1 means that for every y in the range of F, the *upper level set* $L(y) \equiv \{x : F(x) \geq y, x \geq 0_N\}$ is a closed set. The assumption is a technical one of minimal economic interest. It is also a very weak condition from an empirical point of view, since it cannot be contradicted by a finite set of data on the inputs and output of a producer.

If we assume that the production function F satisfies Assumption 1, it turns out that the cost function C has the following properties: *Property 1: C* is a non-negative function; that is, $C(y, p) \geq 0$; *Property 2: C* is *linearly homogeneous* in input prices p for each fixed output level y; that is, $C(y^1, p) \geq C(y^2, p)$ for $y^1 \geq y^2 \geq 0$ and $p \gg 0_N$; *Property 3: C* is *nondecreasing* in p for fixed y; that is,

$C(y, p^1) \geq C(y, p^2)$ for $y \geq 0$, $p^1 \geq p^2 \gg 0_N$; *Property 4*: *C* is *concave* in p for fixed y;

that is, $C(y, \lambda p^1 + (1 - \lambda)p^2) \geq \lambda C(y, p^1) + (1 - \lambda)C(y, p^2)$ for $y \geq 0$, $0 \leq \lambda \leq 1$, $p^1 \gg 0_N$ and $p^2 \gg 0_N$; *Property 5: C* is *nondecreasing* in y for fixed p; that is, *Property 6: C* is *continuous from below* in y for fixed p; that is $\{y : C(y, p) \leq \alpha\}$ is a closed set for every α and $p \gg 0_N$.

Properties 1–4 for C were derived by Shephard (1953) under stronger regularity conditions on F and Properties 4, 5, and 6 were obtained by McKenzie (1957), Uzawa (1964) and Shephard (1970) respectively.

From the viewpoint of economies, all of the properties of C are intuitively obvious except Properties 4 and 6. Property 6 on C is the technical counterpart to Assumption 1 on F and is of minimal economic interest. However, Property 4 has some significant economic implications as we shall see in section "The Comparative Statics Properties of Cost Functions" below.

We can already draw some useful empirical implications from the fact that a cost function must satisfy Properties 1–6 above. For example, in industrial organization and applied econometrics, it is quite common to assume that the true functional form for a firm's or industry's cost function has the following functional form:

$$C(y,p) \equiv \alpha + \beta \cdot p + \gamma y \tag{2}$$

where α , $\beta \equiv (\beta_1, \ \ldots \ , \beta_N)$ and γ are unknown parameters. However, Property 2 implies that α and γ must be zero in order for the cost function to be linearly homogeneous in input prices. But then $C(y, p) = \beta \cdot p$ does not depend on the output level y, which is very implausible.

Duality Between Cost and Production Functions

It is easy to see that the family of upper level sets, $L(y) \equiv \{x : F(x) \geq y, x \geq 0_N\}$, completely determines the production function F. Furthermore, the cost function C may be defined in terms of the production function by (1) or equivalently, in terms of the family of upper level sets as follows:

$$C(y,p) = \min_x \{p \cdot x : x \text{ belongs to } L(y)\}. \tag{3}$$

Thus given the production function F or the family of level sets $L(y)$, the cost function C is determined.

We now ask the following question: given a cost function C which has Properties 1 to 6, can we use C to define the underlying production function F?

For a given output level y and input price vector $p \gg 0_N$, define the corresponding isocost plane by $\{x : p \cdot x = C(y, p)\}$.From the definitions of $C(y, p)$ and $L(y)$, it is obvious that the set $L(y)$ must lie above this isocost plane and be

tangent to it; that is, $L(y)$ must be a subset of the set $\{x : p \cdot x \geq C(y, p)\}$ and this conclusion must be true for every positive input price vector p. Thus $L(y)$ must be a subset of the intersection of all these sets which we denote by $M(y)$:

$$M(y) \equiv \bigcap_{p \gg 0}^{N} \{x : p \cdot x \geq C(y,p)\}. \quad (4)$$

The set $M(y)$ is called the *disposal, convex hull of* $L(y)$; see McFadden (1966).

Each set $\{x : p \cdot x \geq C(y, p)\}$ is a halfspace and is a convex set. A set S *is convex* if and only if x^1 and x^2 belong to S and $0 \leq \lambda \leq 1$ implies $\lambda x^1 + (1 - \lambda)x^2$ also belongs to S. Since $M(y)$ is the intersection of a family of convex sets, $M(y)$ is also a convex set. $M(y)$ also has the following *free disposal* property:

$$x^1 \text{ belongs to } M(y), x^1 \leq x^2, \text{ then } x^2 \text{ belongs to } M(y). \quad (5)$$

We know $L(y)$ must be a subset of $M(y)$. If we want $L(y)$ to coincide with $M(y)$, then $L(y)$ must also be a convex set with the free disposal property. It can be shown that $L(y)$ will have these last two properties for every output level y if and only if the production function F has the following two properties: *Assumption 2 on F: F* is *quasiconcave* function: that is, for every y belonging to the range of F, $L(y) \equiv \{x : F(x) \geq y\}$ is a convex set. *Assumption 3 on F: F* is *nondecreasing;* that is, if $x^2 \geq x^1 \geq 0_N$, then $F(x^2) \geq F(x^1)$.

We may now answer our earlier question about whether a cost function C can completely determine the production function F: the answer is yes if the production satisfies Assumptions 1–3.

More precisely, we have the following result: given a cost function C which satisfies Properties 1–6, then the production function F defined by

$$F(x) \equiv \max_y \{y : C(y,p) \leq p \cdot x \text{ for every } p \gg 0_N\}, x \geq 0_N \quad (6)$$

satisfies Assumptions 1–3. Moreover, if we define the cost function C^* which corresponds to the F defined by (6) in the usual way [recall (1)], then $C^* = C$; that is, this derived cost function C^* coincides with the original cost function C. Thus there is a *duality* between production functions F satisfying Assumptions 1–3 and cost functions C having Properties 1–6: each function completely determines the other under these regularity conditions.

Duality theorems similar to the above results have been established under various regularity conditions by Shephard (1953, 1970), Uzawa (1964), McFadden (1966, 1978a) and Diewert (1971, 1982).

The Derivative Property of the Cost Function

The following result is the basis for most of the theoretical and empirical applications of cost functions.

Suppose the cost function C satisfies Properties 1–6 listed in section "Properties of Cost Functions" and in addition, C is differentiable with respect to the components of p at the point (y^*, p^*). Then the solution $x^* \equiv (x_1^*, \ldots, x_N^*)$ to the cost minimization problem $\min_y \{P^* \cdot x : F(x) \geq y^*, x \geq 0_N\}$ is unique and

$$x_i^* = \partial C(y^*, p^*)/\partial_{pi}, i = 1, \ldots, N; \quad (7)$$

that is, the cost minimizing demand for the ith input is equal to the partial derivative of the cost function with respect to the ith input price.

The result (7) is known as the derivative property of the cost function (see McFadden, 1978a) or Shephard's Lemma, since Shephard (1953) was the first to obtain the result. It should be noted that Hicks (1946) and Samuelson (1947) obtained the result (7) earlier, but under different hypotheses: they assumed the existence of a utility or production function F and deduced (7) by analysing the comparative statics properties of the cost minimization problem (1). On the other hand, Shephard (1953, 1970) assumed only the existence of a cost function satisfying the appropriate regularity conditions.

C

A very elegant proof of (7) using the hypotheses of Hicks and Samuelson is due to Karlin (1959) and Gorman (1976). Their proof proceeds as follows.

Let x^* be a solution to $\min_x\{p^* \cdot x : F(x) \geq y^*, x \geq 0_N\} = C(y^*, p^*)$. Then for every $p \gg 0_N$, x^* is feasible for the cost minimization problem defined by $C(y^*, p) = \min_x\{p \cdot x : F(x) \geq y^*, x \geq 0_N\}$ but it is not necessarily optimal. Thus for every $p \gg 0_N$, we have the following inequality:

$$p \cdot x^* \geq C(y^*, p). \tag{8}$$

We also have

$$p^* \cdot x^* = C(y^*, p^*). \tag{9}$$

For $p \gg 0_N$, define the function $g(p) \equiv p \cdot x^* - C(y^*, p)$. From (8), $g(p) \geq 0$ for all $p \gg 0_N$, and from (9), $g(p^*) = 0$. Thus $g(p)$ attains a global minimum at $p = p^*$. Since g is differentiable, the first-order necessary conditions for a minimum must be satisfied at p^*:

$$\nabla_p g(p^*) = x^* - \nabla_p C(y^*, p^*) = 0_N \tag{10}$$

Where $\nabla_p g(p^*) \equiv [\partial g(p^*)/\partial p_1, \ldots, \partial g(p^*)/\partial p_N]$ denotes the vector of first-order partial derivatives of g with respect to the components of p evaluated at p^* and $\nabla_p C(y^*, p^*)$ denotes the vector of first-order partial derivatives of C with respect to the components of p evaluated at (y^*, p^*). The second set of equalities in (10) can be rearranged to yield (7).

From an econometric point of view, Shephard's Lemma is a very useful result. In order to obtain a valid system of cost minimizing input demand functions, $x(y,p) \equiv [x_1(y,p), \ldots, x_N(y,p)]$ all we have to do is postulate a functional form for C which satisfies Properties 1–6 and then differentiate C with respect to the components of the input price vector p; that is, $x(y, p) = \nabla_p C(y, p)$. It is not necessary to compute the production function F that corresponds to C via the Shephard Duality Theorem nor is it necessary to undertake the often complex algebra involved in deriving the input demand functions using the production function and Lagrangian techniques. In section "Functional Forms for Cost Functions" below, we shall consider several functional forms for C that have been suggested for their econometric convenience.

The Comparative Statics Properties of Cost Functions

Suppose that we are given a cost function C satisfying Properties 1–6 that is also twice continuously differentiable at (y^*, p^*) where $y^* > 0$ and $p^* \gg 0_N$. Applying Shephard's Lemma (7), the above differentiability assumption ensures that the cost minimizing input demand functions $x_i(y, p)$ exist and are once continuously differentiable at (y^*, p^*).

Define $[\partial x_i/\partial p_j] \equiv [\partial x_i(y^*, p^*)/\partial p_j]$ to be the N by N matrix of partial derivatives of the N demand functions $x_i(y^*, p^*)$ with respect to the N prices p_j, $i,j = 1,\ldots, N$. From (7), it follows that

$$\begin{aligned}[\partial x_i/\partial p_j] &= [\partial^2 C(y^*, p^*)/\partial p_i \partial p_j] \\ &\equiv \nabla^2_{pp} C(y^*, p^*)\end{aligned} \tag{11}$$

where $\nabla^2_{pp} C(y^*, p^*)$ is the matrix of second-order partial derivatives of the cost function with respect to the components of the input price vector evaluated at (y^*, p^*). The twice continuous differentiability property of C implies by Young's Theorem in calculus that $\nabla^2_{pp} C(y^*, p^*)$ is asymmetric N by N matrix. Thus using (11), we have

$$[\partial x_i/\partial p_j] = [\partial x_i/\partial p_j]^{\mathrm{T}} = [\partial x_i/\partial p_i] \tag{12}$$

where A^{T} denotes the transpose of the Matrix A. Thus we have established the Hicks (1946) and Samuelson (1947) *symmetry restrictions* on input-demand functions, $\partial x_i(y^*, p^*)/\partial p_j = \partial x_j(y^*, p^*)/\partial p_i$ for all i and j.

Since C is concave in p and is twice continuously differentiable with respect to the components of p at the point (y^*, p^*), it follows from a characterization of concave functions that $\nabla^2_{pp} C(y^*, p^*)$ is a negative semidefinite matrix. Thus by (11),

$$z^{T}[\partial x_i/\partial p_j]z \leq 0 \text{ for all vectors } z. \quad (13)$$

In particular, letting $z = e_i$, the ith unit vector, (13) implies

$$\partial x_i(y^*, p^*)/\partial p_i \leq 0 \text{ for } i = 1, \ldots, N; \quad (14)$$

that is, the ith cost minimizing input demand function cannot slope upwards with respect to the ith input price for $i = 1,..., N$.

Since C is linearly homogeneous in p, we have $C(y^*, \lambda p^*) = \lambda C(y^*, p^*)$ for all $\lambda > 0$. Partially differentiating this equation with respect to p_i for λ close to 1 yields the equation $C_i(y^*, \lambda p^*)\lambda = \lambda C_i(y^*, p^*)$ where $C_i(y^*, p^*) \equiv \partial C(y^*, p^*)/\partial p_i$. Thus $C_i(y^*, \lambda p^*) = C_i(y^*, p^*)$ and differentiation of this last equation with respect to λ yields when $\lambda = 1$:

$$\sum_{j=1}^{N} p_j^* \partial^2 C(y^*, p^*)/\partial p_i \partial p_j = 0 \quad \text{for } i = 1, \ldots, N. \quad (15)$$

Equations (11) and (15) imply that the input-demand functions $x_i(y^*, p^*)$ satisfy the following N restrictions:

$$\sum_{j=1}^{N} p_j^* \partial x_i(y^*, p^*)/\partial p_j = 0 \quad \text{for } i = 1, \ldots, N. \quad (16)$$

A final general restriction on the derivatives of the input-demand functions may be obtained as follows: for λ near 1 differentiate both sides of $C(y^*, \lambda p^*) = \lambda C(y^*, p^*)$ with respect to y and then differentiate the resulting equation with respect to λ. When $\lambda = 1$, the last equation becomes:

$$\sum_{j=1}^{N} p_j^* \partial^2 C(y^*, p^*)/\partial y \partial p_j = \partial C(y^*, p^*)/\partial y. \quad (17)$$

The twice continuous differentiability of C at (y^*, p^*) and (7) imply:

$$\partial^2 C(y^*, p^*)/\partial y \partial p_j = \partial^2 C(y^*, p^*)/\partial p_j \partial y = \partial x_j(y^*, p^*)/\partial y. \quad (18)$$

Property 5 for cost functions implies that

$$\partial C(y^*, p^*)/\partial y^* \geq 0. \quad (19)$$

Using (18) and (19), (17) is equivalent to:

$$\sum_{j=1}^{N} p_j^* \partial x_j(y^*, p^*)/\partial y \geq 0. \quad (20)$$

Thus for at least one j, we must have $\partial x_j(y^*, p^*)/\partial y \geq 0$; that is, as output increases, not every input demand can decrease.

We have shown that the assumption of cost minimizing behaviour implies a number of restrictions on input demand functions that are potentially testable. Hicks (1946) and Samuelson (1947) obtained the restrictions (12), (13), and (16) using the first-order conditions for the primal cost minimization problem (1) and the properties of determinants of bordered Hessian matrices; Samuelson also obtained (20). Our derivation of the restrictions on input-demand functions using the dual approach is due to McKenzie (1957), Karlin (1959) and McFadden (1978a).

Hicks (1946) also showed that when $N = 2$, so that there are only two inputs, then (12), (13), and (16) imply that

$$\partial x_1(y^*, p_1^*, p_2^*)/\partial p_2 = \partial x_2(y^*, p_1^*, p_2^*)/\partial p_1 \geq 0. \quad (21)$$

Hicks (1946) called two distinct goods i and j *substitutes* if and only if $\partial x_i(y, p)/\partial p_j \geq 0$ and *complements* if and only if $\partial x_i(y, p)/\partial p_j < 0$. Thus in the two input case, the two goods must be substitutes. Hicks also showed that in the three input case, at least two of the three pairs of goods must be substitutes.

We turn now to empirical applications of cost functions.

C

Functional Forms for Cost Functions

Shephard's Lemma (7) provides a convenient method for generating systems of cost minimizing input demand functions: simply postulate a functional form for $C(y, p)$ and then partially differentiate C with respect to each input price. Below, we present three examples to illustrate the technique.

Our first example is the *translog cost function* due to Christensen et al. (1971, 1973). The logarithm of the cost function is defined as follows:

$$\begin{aligned}\ln C(y,p) &\equiv \alpha_0 + \sum_{i=1}^{N} a_i \ln p_i \\ &= (1/2) \times \sum_{i=1}^{N}\sum_{j=1}^{N} a_{ij} \ln p_i \ln p_j \\ &\quad + \sum_{i=1}^{N} a_{ij} \ln p_i \ln y + a_y \ln y \\ &\quad + (1/2) a_{yy} \ln y \ln y \end{aligned} \tag{22}$$

where the a_i, a_{ij}, $= a_{ij}$, a_i, a_y and a_{yy} are $1 + N + (1/2)N \quad (N + 1) + N + 2 = 3 + 2N +$ $(1/2)N(N + 1)$ parameters determined by the technology of the firm or industry. Differentiating both sides of (22) with respect to the logarithm of the ith input price, $\ln p_i$ for $i = 1, \ldots, N$ yields the following system of equations:

$$s_i = a_i + \sum_{j=1}^{N} a_{ij} \ln p_j + a_{iy} \ln y, \quad i = 1, \ldots, N \tag{23}$$

where the ith input cost share is defined as $s_i \equiv [p_i \partial C(y, p)/\partial p_i]/C(y, p) = p_i x_i(y, p)/C(y, p)$ where the last equality follows using (7).

By Property 2 for cost functions, $C(y, p)$ must be linearly homogeneous in input prices. This property will be satisfied by the translog cost function cost function if and only if the following N+2 linear restrictions on the parameters hold:

$$\sum_{i=1}^{N} a_i = 1, \sum_{i=1}^{N} a_{iy} = 0 \text{ and } \sum_{j=1}^{N}$$

$$a_{ij} = 0 \text{ for } i = 1, \ldots, N. \tag{24}$$

It is possible to append errors to equations (22) and N–1 of the equations (23) and econometrically estimate the unknown parameters, given data on inputs, input prices and output. The symmetry restrictions $a_{ij} = a_{ij}$ and the restrictions (24) may be imposed or one can test for their validity. If these restrictions are imposed, then the resulting translog cost function will have $1 + N + (1/2)$ $N(N + 1)$ free parameters.

What considerations are relevant in choosing a functional form for a cost function? The following four properties are desirable: (i) *flexibility*; that is, the functional form for C should have a sufficient number of free parameters to be able to provide a second-order approximation to an arbitrary twice continuously differentiable function with the appropriate theoretical properties, (ii) *parsimony*; that is, the functional form for C should have the minimal number of free parameters required to have the flexibility property, (iii) *linearity*; that is, the unknown parameters of C should appear in the system of estimating equations in a linear fashion in order to facilitate econometric estimation, and (iv) *consistency*; that is, the functional form for C should be consistent with Properties 1–6 for cost functions. These considerations were first suggested by Diewert (1971) in an informal manner; the term parsimony is due to Fuss et al. (1978) and the term flexible is due to Diewert (1974). The equivalence of various definitions of the flexibility property is discussed by Barnett (1983).

How satisfactory is the translog cost function in the light of the above considerations? We consider the flexibility property first. In order to be able to approximate a function of $1 + N$ variables to the second order, we require $1 + (1 + N) + (1 + N)^2$ free parameters. However, if we assume that the cost functions are twice continuously differentiable, then we can reduce the number by $N(N + 1)/2$ due to the symmetry property of the second-order partial derivatives. The linear homogeneity property of the cost function, Property 2, yields an additional N+1 restriction on the first and second derivatives of C, (15) and (17), plus the

following restriction (which follows using Euler's Theorem on homogeneous functions):

$$C(y,p) = \sum_{i=1}^{N} p_i \partial C(y,p)/\partial p_i. \qquad (25)$$

Thus a flexible functional form for a cost function should have $1 + (1 + N) + (1 + N)^2 - [(1/2)N(N + 1) + N + 1 + 1] = 1 + N + (1/2)N(N + 1)$ free parameters, which is precisely the number the translog cost function has when the restrictions (24) are imposed. It can be shown that the translog cost function is indeed flexible and we have just shown that it is also parsimonious.

As can be seen by inspecting (22) and (23), the estimating equations are linear in the unknown parameters, so the linearity property is also satisfied.

If the restrictions (24) are imposed, Property 2 will be satisfied. In practice, the other properties that a cost function must have will be satisfied with the exception of Property 4, the concavity in prices property. If all of the a_{ij} and a_{iy} parameters are zero, then the translog cost function reduces to a Cobb–Douglas cost function which satisfies the concavity property globally. However, in the general case, the best we can hope for is that the concavity property is satisfied locally for a range of input prices.

If a production function is linearly homogeneous (that is, $F(\lambda x) = \lambda F(x)$ for $\lambda \geq 0$ and $x \geq 0_N$) so that the technology is subject to constant returns to scale, then the corresponding cost function has the following property:

$$C(y,p) = yC(1,p); \qquad (26)$$

that is, total cost is equal to the output level y times the cost of producing one unit of output, $C(1, p) \equiv c(p)$, the *unit cost function.*

If C is twice continuously differentiable and satisfies (26), then one can show that the following $2 + N$ restrictions on the first and second derivatives of C must hold:

$$C(y,p) = y\partial C(y,p)\partial y; \qquad (27)$$

$$\partial^2 C(y,p)/\partial y^2 = 0; \qquad (28)$$

$$\partial C(y,p)/\partial p_i = y\partial^2 C(y,p)/\partial y \partial p_i, \quad i = 1, \ldots, N. \qquad (29)$$

However, in view of (25), it can be seen that only N–1 of the restrictions (29) are new. Thus the assumption of a constant returns to scale technology imposes new restrictions on the derivatives of the cost function C.

It can be shown that necessary and sufficient conditions for the translog cost function defined by (22) and (24) to satisfy (26) are the following N+1 restrictions:

$$a_y = 1, a_{yy} = 1 \text{ and } a_{iy} = 0 \text{ for } i = 1, \ldots, N-1. \qquad (30)$$

Of course (30) and (24) imply that a_{Ny}=0 as well.

It can be shown that if the restrictions (24) and (30) are imposed on the parameters of the translog cost function defined by (22), then the resulting functional form is flexible in the class of cost functions that satisfy the constant returns to scale property (26). Note that we can test for the validity of the constant returns to scale property by testing whether the N+1 linear restrictions (30) hold.

For our second example, consider the following functional form for a cost function:

$$C(y,p) \equiv c(p)y + \sum_{i=1}^{N} b_i p_i + b_{yy}\left(\sum_{i=1}^{N} \beta_i p_i\right) y^2; \qquad (31)$$

$$c(p) \equiv \sum_{i=1}^{N}\sum_{j=1}^{N} b_{ij} p_i^{1/2} p_j^{1/2} \qquad (32)$$

where the b_{yy}, b_i, $b_{ij} = b_{ji}$ and β_i are parameters which characterize the technology. If $b_i = 0$ for $i = 1, \ldots, N$ and $b_{yy} = 0$, then (31) reduces to the *Generalized Leontief cost function* defined by Diewert (1971). If in addition, $b_{ij} = 0$ for all $i \neq j$, then (31) reduces to the cost function $\sum_{i=1}^{N}$

$b_{ii}p_iy$, which is dual to the Leontief (no substitution) production function, $F(x_i, \ldots, x_N) \equiv \min \{x_i/b_{ii} : i = 1, \ldots, N\}$.

In order for the cost function defined by (31) and (32) to satisfy the parsimony property, it is necessary for the empirical investigator to prespecify the β_i parameters; for example, one could set β_i equal to 1 or to the average input quantity x_i observed in the sample of data. Under these conditions, the Generalized Leontief cost function has $(1/2)N(N + 1) + N + 1$ free parameters, which is just the required number for the flexibility property. In fact, Diewert and Wales (1987) show that this cost function is flexible and parsimonious when the β_i are predetermined.

Applying (7), the input-demand functions that correspond to (31) and (32) are:

$$x_i(y,p) = \sum_{j=1}^{N} b_{ij}p_i^{-1/2}p_j^{1/2}y + b_i + b_{yy}\beta_i y^2.i = 1, \ldots, N \tag{33}$$

For the purpose of econometric estimation, errors can be appended to the N Eq. 33. If the βi are predetermined, it can be seen that the system of estimating equations is linear in the unknown parameters.

If we wish to test for a constant returns to scale technology, then the following $1+N$ linear restrictions on the parameters are necessary and sufficient for this property:

$$b_{yy} = 0 \text{ and } b_i = 0 \text{ for } i = 1, \ldots, N. \tag{34}$$

Note that the linear homogeneity in prices property is satisfied by the Generalized Leontief cost function. The other properties for cost functions will also be satisfied in practice with the exception of Property 4, the concavity in prices property. If all $b_{ij} \geq 0$ for $i \neq j$, then the concavity property will be globally satisfied, but this assumption rules out complementary pairs of inputs (recall the discussion about substitutes and complements at the end of the previous section). Thus in general, one can only hope that the concavity property will be satisfied locally, as was the case with the translog cost function.

For our third and final example, consider the following *normalized quadratic cost function* defined by (31) but now $c(p)$ is defined as follows:

$$c(p) \equiv \sum_{i=1}^{N} b_{ii}p_i + (1/2)\sum_{i=1}^{N} \times \sum_{j=1}^{N} a_{ij}p_ip_j/\left(\sum_{n=1}^{N}\alpha_n p_n\right) \tag{35}$$

where the N by N matrix $A \equiv [a_{ij}]$ is symmetric and satisfies the following restriction for some input price vector $p^* \gg 0_N$;

$$Ap^* = 0_N. \tag{36}$$

This functional form is due to Diewert and Wales (1987); its generalizes some functional forms due to Fuss (1977) and McFadden (1978b). The functional form has Nb_{ii} parameters, $(1/2)N(N - 1)$ free a_{ij} parameters taking into consideration (36), N b_i parameters, 1 b_{yy}, N β_i and $N\alpha_n$ parameters or $1 + 3N + (1/2)N(N + 1)$ free parameters in all. In order for this functional form to have the parsimony property, it is necessary for the empirical investigator to prespecify the β_i and α_n parameters; we assume that this has been done and these parameters are non-negative and not identically equal to zero. Under these conditions, Diewert and Wales (1987) show that this cost function is parsimonious and flexible at the point (y^*, p^*) where p^* is the price vector which appears in (36).

Applying (7), the system of input demand functions divided by the output y is:

$$x_i(y,p)/y = b_{ii} + \sum_{j=1}^{N} a_{ij}p_j\left(\sum_{n=1}^{N}\alpha_n p_n\right)^{-1} - \left(\sum_{j=1}^{N}\sum_{k=1}^{N} a_{jk}p_jp_k\right) \times \left(\sum_{n=1}^{N}\alpha_n p_n\right)^{-2}\alpha_i + b_i y^{-1} + b_{yy}\beta_i y, i = 1, \ldots, N. \tag{37}$$

Errors can be appended to (37) and we obtain a system of estimating equations which is linear in

the unknown parameters, provided that the α_n and β_i are prespecified.

If we wish to test for a constant returns to scale technology, then again the $1 + N$ linear restrictions (34) are necessary and sufficient for this property.

The normalized quadratic cost function with prespecified α_n and β_i is flexible, parsimonious and has linear estimating equations. As was the case with our first two examples, our third example has no problem in satisfying Properties 1, 2, 3, 5 and 6 for cost functions. It also turns out that our third example has no problem in satisfying Property 4: Diewert and Wales (1986) using some results due to Lau (1978), show that the normalized quadratic cost function is globally concave if and only if the A matrix is negative semidefinite. They also indicate how this negative semidefiniteness property can be imposed if necessary without destroying the flexibility of the functional form; simply set $A = -SS^{T}$ where S is a lower triangular N by N matrix and S^{T} is its transpose. However, in this latter case, nonlinear regression techniques must be used in order to estimate the unknown parameters.

The extensive empirical literature on estimating cost functions is nicely reviewed by Jorgenson (1984).

Applications to the Estimation of Consumer Preferences

The cost function techniques described in the previous section can be used to obtain empirical descriptions of technologies. Those techniques can also be adapted to obtain empirical descriptions of consumer preferences.

As was noted in section "Properties of Cost Functions," y may be interpreted as a household's welfare level, F as a utility or preference function, p as a vector of commodity prices and $C(y, p)$ as the minimum cost of achieving at least the welfare level y.

However, the econometric techniques described in the previous section cannot be utilized immediately in the consumer context because utility cannot be observed whereas output can. We acknowledge this difference by using u, the consumer's utility or welfare level, in place of y in what follows.

The theory outlined in the previous sections is still valid: given a differentiable functional form for the cost function $C(u, p)$ that satisfies Properties 1 to 6, we may form the consumer's system of *constant real income or Hicksian demand functions* $x(u, p) \equiv [x_1(u, p), \ldots, x_N(u, p)]$ by differentiating the cost function with respect to each commodity price p_i [recall (7)]:

$$x_i(u,p) = \partial C(u,p)/\partial p_i, i = 1, \ldots, N. \quad (38)$$

We determine u as a function of the prices p and the consumer's observed expenditure on commodities during the period Y, say, by equating the minimum cost of achieving the welfare level u to the observed expenditure; that is, we solve the following equation for u:

$$C(u,p) = Y. \quad (39)$$

The solution function g where $u = g(Y, p)$ is known as the consumer's *indirect utility function.* Now replace u in the right-hand side of (38) by $g(Y, p)$ and obtain the consumer's system of *market demand functions:*

$$x_i = \partial C(g(Y,p),p)/\partial p_i, i = 1, \ldots, N. \quad (40)$$

If we multiply equation i in (40) by p_i, sum the resulting equations and use (7), (25) and (39), then we obtain the identity $\sum_{i=1}^{N} p_i x_i = Y$, so only N–1 of the N equations in (40) are independent. Thus for econometric estimation purposes, we may add errors to N–1 of the equations in (40), and given a functional form for C, we may use these equations to estimate the unknown parameters in C. We shall discuss this technique in more detail shortly, but first, we must discuss the problems involved in cardinalizing utility.

The scaling of utility is irrelevant in describing a consumer's preferences. However, when we postulate a functional form for a cost function, we are implicitly imposing a cardinalization of the consumer's utility. Hence, we might as well

impose a convenient cardinalization: *money metric scaling of utility* (the term is due to Samuelson 1974). This involves setting utility u equal to 'income' Y, holding prices constant at some specified price vector p^*, that is, we have

$$Y = g(Y, p^*) \quad \text{for all } Y > 0. \tag{41}$$

In terms of the cost function, (41) may be rewritten as

$$u = C(u, p^*) \text{ for all } u > 0. \tag{42}$$

In examples 2 and 3 in the previous section, the cost function had the following form:

$$C(u,p) = c(p)u + \sum_{i=1}^{N} b_i p_i + b_{yy}\left(\sum_{i=1}^{N} \beta_i p_i\right) u^2. \tag{43}$$

In order to make (43) consistent with money metric scaling, (42), the following three restrictions on the parameters of C must be satisfied:

$$c(p^*) = 1, \sum_{i=1}^{N} b_i p_i^* = 0 \text{ and } b_{yy} = 0. \tag{44}$$

Using $b_{yy} = 0$ we find that the indirect utility function that corresponds to the C defined by (43) is

$$g(Y,p) = \left(Y - \sum_{i=1}^{N} b_i p_i\right)/c(p). \tag{45}$$

Substitution of (45) into (40) yields the following system of consumer demand functions:

$$x_i = b_i + [\partial c(p)/\partial p_i]\left(Y - \sum_{i=1}^{N} b_i p_i\right)/c(p). \quad i = 1, \ldots, N. \tag{46}$$

Now add errors to N–1 of the Eq. 46, calculate the partial derivatives of the $c(p)$ defined by (32) or (35), impose the normalizations (44) and we have a system of nonlinear estimating equations. An empirical example of this technique for estimating consumer preferences may be found in Diewert and Wales (1986).

Finally, we note that cost functions of the type defined by (43) with $b_{yy} = 0$ have very convenient aggregation over consumers' properties; see Gorman (1953) and Deaton and Muellbauer (1980).

Cost Functions and Measures of Welfare Gain

Consider a consumer whose preferences can be represented by the differentiable cost function, $C(u, p)$. Suppose we can observe the consumer's choices x^1 and x^2 during periods 1 and 2 when prices p^1 and p^2 prevail. Let u^1 and u^2 be the welfare levels attained during those two periods. Then by (7),

$$x^i = \nabla_p C(u^i, p^i), i = 1, 2. \tag{47}$$

For many purposes in applied welfare economics, it is useful to evaluate the *ex post* welfare change of the consumer. Two natural measures, suggested originally by Hicks (1942), are his *equivalent* and *compensating variations* which we denote by $V(p^1)$ and $V(p^2)$:

$$V(p^1) \equiv C(u^2, p^1) - C(u^1, p^1); V(p^2) \equiv C(u^2, p^2) - C(u^1, p^2). \tag{48}$$

From (47) and (25), $C(u^1, p^1) = p^1 \cdot x^1$ and $C(u^2, p^2) = p^2 \cdot x^2$. However, the costs $C(u^2, p^1)$ and $C(u^1, p^2)$ are not observable. Hence the following question arises: can we form approximations to $V(p^i)$ that use only observable data?

Linear approximations to $C(u^i, p^j)$ may be obtained using Taylor's Theorem. Thus we have:

$$V(p^1) \cong [C(u^2, p^2) + \nabla_p C(u^2, p^2) \cdot (p^1 - p^2)] - C(u^1, p^1) = [p^2 \cdot x^2 + x^2 \cdot (p^1 - p^2)] - p^1 \cdot x^1 \text{ using (47)} = p^1 \cdot (x^2 - x^1) \tag{49}$$

and

$$
\begin{aligned}
V(p^2) &\cong C(u^2,p^2) \\
&- \left[C(u^1,p^1) + \nabla_p C(u^1,p^1)\cdot(p^2-p^1)\right] \\
&= p^2\cdot x^2 - \left[p^1\cdot x^2 + x^1\cdot(p^2-p^1)\right] \text{ using (47)} \\
&= p^2\cdot(x^2-x^1).
\end{aligned}
\tag{50}
$$

The first-order approximations (49) and (50) are essentially due to Hicks (1942, 1946).

To obtain a second-order approximation result, we proceed indirectly. Suppose the consumer's cost function is defined by

$$
C(u,p) \equiv c(p) + \sum_{i=1}^{N} b_i p_i u \tag{51}
$$

where $c(p)$ is the normalized quadrative unit cost function defined by (35) for some prespecified $\alpha \equiv (\alpha_1, \ldots, \alpha_n) > 0_N$.

It can be shown that the cost function defined by (51) can provide a second- order approximation to an arbitrary twice continuously differentiable cost function that satisfies the money metric scaling of utility property (42).

Now use the parameters vector α which occurred in the definition of $c(p)$ in order to define the *normalized prices* v^i

$$
v^i \equiv p^i/(p^i\cdot\alpha), \quad i=1,2. \tag{52}
$$

Straightforward calculations show that if C is defined by (51), then the following identity holds *exactly*:

$$
\begin{aligned}
&(1/2)V(v^1) + (1/2)V(v^2) \\
&\quad = (1/2)(v^1+v^2)\cdot(x^2-x^1)
\end{aligned}
\tag{53}
$$

where $V(v^1)$ and $V(v^2)$ are equivalent and compensating variations evaluated using the normalized prices v^i in place of the commodity price vectors p^i. Thus (53) says that an average of the Hicksian variations using normalized prices is exactly equal to the average of the normalized prices inner producted with the vector of quantity differences, x^2–x^1, provided that preferences are defined by the cost function (51). Note that the right-hand side of (53) can be evaluated using observable price and quantity data. Since the formula on the right-hand side is exact for preferences which have a second-order approximation property, we could call it a *superlative welfare gain measure* in analogy to the terminology used in index number theory. The term gain measure is due to King (1983).

See Also

- Duality
- Production Functions

Bibliography

Barnett, W.A. 1983. New indices of money supply and the flexible Laurent demand system. *Journal of Business and Economic Statistics* 1: 7–23.

Christensen, L.R., D.W. Jorgenson, and L.J. Lau. 1971. Conjugate duality and the transcendental logarithmic production function. *Econometrica* 39: 255–256.

Christensen, L.R., D.W. Jorgenson, and L.J. Lau. 1973. Transcendental logarithmic production frontiers. *The Review of Economics and Statistics* 55: 28–45.

Deaton, A., and J. Muellbauer. 1980. *Economics and consumer behaviour*. Cambridge: Cambridge University Press.

Diewert, W.E. 1971. An application of the Shephard duality theorem: A generalized Leontief production function. *Journal of Political Economy* 79: 481–507.

Diewert, W.E. 1974. Applications of duality theory. In *Frontiers of quantitative economics*, ed. M.D. Intriligator and D.A. Kendrick, vol. 2, 106–171. Amsterdam: North-Holland.

Diewert, W.E. 1982. Duality approaches to microeconomic theory. In *Handbook of mathematical economics*, ed. K.J. Arrow and M.D. Intriligator, vol. 2. Amsterdam: North-Holland.

Diewert, W.E., and T.J. Wales. 1986. Normalized quadratic systems of consumer demand functions. Discussion Paper No. 86–16, Department of Economics, University of British Columbia, Vancouver, May.

Diewert, W.E., and T.J. Wales. 1987. Flexible functional forms and global curvature conditions. *Econometrica* 55: 43–68.

Fuss, M.A. 1977. Dynamic factor demand systems with explicit costs of adjustment. In *Dynamic models of the industrial demand for energy*, ed. E.R. Berndt, M. Fuss, and L. Waverman. Palo Alto: Electric Power Research Institute.

Fuss, M., D. McFadden, and Y. Mundlak. 1978. A survey of functional forms in the economic analysis of production. In *Production economics: A dual approach to theory and applications*, ed. M. Fuss and D. McFadden, vol. 1. Amsterdam: North-Holland.
Gorman, W.M. 1953. Community preference fields. *Econometrica* 21: 63–80.
Gorman, W.M. 1976. Tricks with utility functions. In *Essays in economic analysis*, ed. M. Artis and R. Nobay. Cambridge: Cambridge University Press.
Hicks, J.R. 1942. Consumers' surplus and index-numbers. *Review of Economic Studies* 9: 126–137.
Hicks, J.R. 1946. *Value and capital*. 2nd ed. Oxford: Clarendon Press.
Jorgenson, D.W. 1984. Econometric methods for modeling producer behaviour. Discussion Paper No. 1086, Harvard Institute for Economic Research, Harvard University, Cambridge, MA. In *Handbook of econometrics*, vol. 3, ed. Z. Griliches and M.D. Intriligator. Amsterdam: North-Holland.
Karlin, S. 1959. *Mathematical methods and theory in games, programming and economics*. Vol. 1. Palo Alto: Addison-Wesley.
King, M.A. 1983. Welfare analysis of tax returns using household data. *Journal of Public Economics* 21: 183–214.
Lau, L.J. 1978. Testing and imposing monotonicity, convexity and quasi-convexity constraints. In *Production economics: A dual approach to theory and applications*, ed. M. Fuss and D. McFadden, vol. 1. Amsterdam: North-Holland.
McFadden, D. 1966. *Cost, revenue and profit functions: A cursory review*. Working Paper No. 86, IBER, University of California at Berkeley, March.
McFadden, D. 1978a. Cost, revenue and profit functions. In *Production economics: A dual approach to theory and applications*, ed. M. Fuss and D. McFadden, vol. 1. Amsterdam: North-Holland.
McFadden, D. 1978b. The general linear profit function. In *Production economics: A dual approach to theory and applications*, ed. M. Fuss and D. McFadden, vol. 1. Amsterdam: North-Holland.
McKenzie, L. 1957. Demand theory without a utility index. *Review of Economic Studies* 24: 185–189.
Samuelson, P.A. 1947. *Foundations of economic analysis*. Cambridge: Harvard University Press.
Samuelson, P.A. 1974. Complementarity – an essay on the 40th anniversary of the Hicks–Allen revolution in demand theory. *Journal of Economic Literature* 12: 1255–1289.
Shephard, R.W. 1953. *Cost and production functions*. Princeton: Princeton University Press.
Shephard, R.W. 1970. *Theory of cost and production functions*. Princeton: Princeton University Press.
Uzawa, H. 1964. Duality principles in the theory of cost and production. *International Economic Review* 5: 216–220.

Cost Minimization and Utility Maximization

Peter Newman

Keywords

Arrow corner; Compensated demand; Cost function; Cost minimization and utility; Duality; Indirect utility function; Linear programming; Maximization

JEL Classifications

D1

Consider the following standard problem in the theory of demand: Find $x^* \geq 0$ so as to max $u(x)$ subject to $\langle x, p\rangle \leq \omega$ where $\langle x, p\rangle$ is the inner product of the n-dimensional commodity and price vectors, and $\omega > 0$ and u are the consumer's income and utility function respectively; this problem is here labelled $max(p, \omega)$.

The functional dependence of the *value* $v^*[\equiv u(x^*)]$ of this nonlinear programming problem on its parameters (p, ω) is denoted by $v(p, \omega)$, where v is the *indirect* utility function. The similar dependence of the *solution* x^* of max (p, ω) is written $f(p, \omega)$, where f is the *ordinary* (or *Marshallian*) demand function (or correspondence). If v^* does not exist then neither do v, x^* or f. Important though they are such non-existence problems are irrelevant here, so without further ado assume that every optimization problem has a solution.

Consider next a problem whose form is similar to that of max (p, ω) but whose objective is different, that is, cost minimization rather than utility maximization. Specifically, find $x^{**} \geq 0$ so as to minmin $\langle x, p\rangle$ subject to $u(x) \geq \tau$ where x, p and u are as before and τ is a *target* level of utility; this new problem is labelled $\min(p, \tau)$. The functional dependence of the value $\mu^{**}(\equiv\langle x^{**}, p\rangle)$ of $\min(p, \tau)$ on its parameters (p, τ) is denoted $c(p,$

τ), where c is the *cost* (or *expenditure*) function. The similar dependence of x^{**} on (p, τ) is written $h(p, \tau)$, where h is the *compensated* (or *Hicksian*) demand function (or correspondence).

Suppose now that $\max(p, \omega)$ is solved and its value v^* is inserted into the second optimization problem, thus creating the problem $\min(p, v^*)$. Is each solution x^* of $\max(p, \omega)$ necessarily also a solution of $\min(p, v^*)$? Call this Question I. A similar question can be asked of the reverse situation, which is: For arbitrary (p, τ) solve $\min(p, \tau)$, obtain it value μ^{**} and then solve the resulting max problem, (p, μ^{**}). Question II is then: Is each solution x^{**} of $\min(p, \tau)$ necessarily also a solution of $\max(p, \mu^{**})$?

Problem $\min(p, v^*)$ has often been called the *dual* of $\max(p, \omega)$, from as far back as Arrow and Debreu (1954, pp. 285–6) to Deaton and Muellbauer (1980, pp. 37 ff.) and beyond. Indeed, this usage is now so common that for most economists $\min(p, v^*)$ seems to be *the* leading species of the genus *dual problem*.

One can see why. It appears to be quite analogous to dual problems in linear programming (lp), with max becoming min, and objective and constraint functions becoming constraint and objective functions, respectively. However, the analogy with duals in lp is misleading, for each solution x^{**} of the alleged 'dual' $\min(p, v^*)$ is located in the *same* space as each solution x^* of its 'primal' $\max(p, \omega)$, whereas in lp the solutions to the dual all lie in the *dual* space. As Deaton and Muellbauer justly remark: 'The essential feature of the duality approach is a *change of variables*' (1980, p. 47, their italics). So a new term for the relation that $\min(p, v^*)$ bears to $\max(p, \omega)$ is needed in order to distinguish it from genuine duality; the 'mirrored' (or 'reflected') problem is suggested in Newman (1982).

In demand theory it is sometimes recognized explicitly that Question I needs an answer (for example, Samuelson 1947, p. 103; McKenzie 1957, p. 186) but more often not, probably because the usual assumptions on preferences are quite sufficient for coincidence of x^{**} with x^*. An explicit treatment appears unnecessary: '... clearly, the vector of commodities must in both cases be the same' (Deaton and Muellbauer 1980, p. 37). In welfare economics, however, it has long been recognized that a suitably generalized form of Question I, simple as it is, has importance for the first fundamental theorem of welfare economics, namely that every competitive allocation is (strongly) Pareto-optimal.

Question II has always been considered more delicate than Question I. Indeed, it was not even put until Arrow (1951, pp. 527–8) exhibited his famous 'exceptional case' (now often known as the Arrow Corner) in which it receives a negative answer. Its relevance for proofs of existence of competitive equilibrium was fully grasped by Arrow and Debreu (1954, sections 4 and 5), and later Debreu (1959, pp. 67–71), for essentially this reason, devoted four pages of his terse classic to a detailed examination of both Questions.

It is interesting that although the second Question is economically more subtle than the first, from a sufficiently abstract point of view the two are logically isomorphic (see Newman 1982, where in both Theorem (c) and Theorem (c′) the assertion 'iff' is wrong and should be replaced by 'if'). While such extreme abstraction is irrelevant here, both $\max(p, \omega)$ and $\min(p, \tau)$ do need to be put into a form suitable for general equilibrium theory.

The Setting

The consumer is now endowed, not with an exogenous positive income, but with a nonzero bundle x^0, whose worth $\langle x^0, p\rangle$ may be zero. For simplicity (and only that), free disposal is assumed.

Assumptions About Preferences

The consumer has two disjoint binary relations $\succ$ ('preference') and $\sim$ ('indifference') each defined on some non-empty $S \subset R^n$; the union of $\succ$ and $\sim$ is denoted $\succsim$. Indifference is reflexive and symmetric (so that preference is irreflexive) and the

statements $x^1 \succ x^3$ and $x^2 \sim x^3$ together imply $x^1 \succ x^3$. Neither completeness nor transitivity of preference is assumed, so a utility function need not exist.

The generalized version of max(p, ω) is then: Find $x^* \in S$ for which $\langle x^{**} p\rangle \leq \langle x^0, p\rangle$ and such that $x \succ x^*$ implies $\langle x, p\rangle > \langle x^0, p\rangle$. In words, '$x^*$ is feasible and anything preferred to it is unaffordable'. This problem is labelled max(p, x^0).

The generalized version of min(p, τ) is: Find $x^{**} \in S$ for which $x^{**} \gtrsim t$ and such that $x \gtrsim t$ implies $\langle x, p\rangle \geq \langle x^{**}, p\rangle$. In words, 'anything at least as good as the target bundle $t \in S$ costs at least as much as x^{**}'. This problem is labelled min(p, t).

Note that in the absence of a utility function max(p, x^0) can have a solution but not a value, while min(p, t) can have both value and solution, just as before.

Some Definitions

Any bundle $x^{\#}$ to which no $x \in S$ is preferred is called *bliss*, while a bundle $x_{\#}$ for which at prices p there is no cheaper $x \in S$ is called *p-minimal*. Preferences are *locally nonsatiated* at x^1 if *any* neighbourhood $N(x^1)$ contains $x \succ x^1$, while $x^2 \in S$ has *locally cheaper points* at p (a term apparently due to McKenzie 1957) if *any* neighbourhood $N(x^2)$ contains a bundle $x \in S$ which at prices p is cheaper than x^2. If $x^{\#}$ is bliss it cannot be locally nonsatiated, and if $x_{\#}$. is p-minimal it has no locally cheaper points at p.

Following Bergstrom et al. (1976), preferences are said to have *open upper sections* if $x' \succ x^2$ implies the existence of a neighbourhood $N(x^1) \subset S$ for which $x \succ x^2$ for every x in it.

The following simple result answers both Questions satisfactorily and generalizes easily to a wide class of infinite-dimensional commodity spaces.

Theorem

(i) Assume (a) that if $x \in S$ is not bliss it is locally nonsatiated, and (b) that the solution x^* of max(p, x^0) is not bliss. Then x^* also solves min(p, x^*). Moreover, the value μ^{**} of min(p, x^{**}) equals $\langle x^0, p\rangle$.

(ii) Assume (c) that preferences have open upper sections, (d) that if $x \in S$ is not p-minimal it has locally cheaper points at p, and (e) that the solution x^{**} of min(p, t) is not p-minimal. Then x^{**} also solves max(p, μ^{**}), where $\mu^{**} = \langle x^{**}, p\rangle$.

Proof

(i) Suppose the result false, so there exists $x^1 \gtrsim x^*$ such that $\langle x^1, p\rangle < \langle x^*, p\rangle$ Now $x^1 \gtrsim x^*$ cannot occur because if it did $\langle x^1, p\rangle < \langle x^*, p\rangle \leq \langle x^0, p\rangle$ would imply that x^* does not solve max(p, x^0), contrary to hypothesis. So $x^1 \sim x^*$.

Since the vector p represents a continuous linear function (al) there is a neighbourhood $N(x^1)$ all of whose points are cheaper at prices p than x^*. From (b) there exists $x \succ x^*$, and this and the symmetry of $\sim$ imply that $x \succ x^1$ as well, so that x^1, is not bliss either. Hence from (a) at least one member of $N(x^1)$, say x^2, is such that $x^2 \succ x^1$. Because $x^1 \succ x^*$ this leads to $x^2 \succ x^*$, which again contradicts the hypothesis. Thus x^* solves min(p, x^*), which implies $\mu^{**} = \langle x^*, p\rangle$.

Suppose $\langle x^*, p\rangle < \langle x^0, p\rangle$. By the continuity of p there exists $N(x^*)$ all of whose points are cheaper at prices p than is x^0, while from (b) and (a) at least one of them, say x^3, is such that $x^3 \succ x^*$. Yet again, this contradicts the hypothesis. So $\langle x^*, p\rangle = \langle x^0, p\rangle$.

(ii) By assumption $x^{**} \gtrsim t$ Suppose $x^{**} \succ t$. From (c) there exists $N(x^{**})$ such that $x \succ t$ for every x in it. From (e) and (d) x^{**} has locally cheaper points at p, so at least one x in $N(x^{**})$, say x^1, is cheaper than x^{**} at p. Since $x^1 \succ t$, this contradicts the hypothesis that x^{**} solves min(p, t). Hence $x^{**} \sim t$.

Suppose now that x^{**} does not solve max(p, μ^{**}), so there is an x^2 such that $x^2 \succ x^{**}$ and $\langle x^2, p\rangle \leq \langle x^{**}, p\rangle$. Hence $x^2 \succ t$. If x^2 were cheaper at p than x^{**} that would again contradict the hypothesis. So $\langle x^2, p\rangle \leq \langle x^{**}, p\rangle$.

From (e) there is an $x \in S$ cheaper at p than x^{**}, hence cheaper than x^2, so x^2 is not p-minimal either. Since $x^2 \succ t$, from (c) there exists $N(x^2)$ such that $x \succ t$ for every x in it and from (d) at least one of these must be cheaper than x^2 at p, and so cheaper than x^{**}, which again contradicts the hypothesis. Q.E.D.

One sees just how few and how weak are the assumptions on preferences that enable Questions I and II to be answered, as distinct (for ex) from those needed to guarantee the *existence* of solutions x^* and x^{**}. Note that two assumptions are used for Question I and three for II, an inequality which occurs because the constraint in $\max(p, x^0)$ is linear and hence continuous, whereas in the problem $\min(p, t)$ some continuity in the (nonlinear) constraint has to be *imposed* by means of the 'extra' assumption (c). This asymmetry disappears in a more abstract treatment, with more general constraints.

The intuitions behind the proof help to see why Question II is a serious problem for general equilibrium theory. In the proof of (i) the bundle x^1 that is cheaper than x^* is made a little bigger, in effect increasing satisfaction by increasing expenditure, until a bundle is reached that is still affordable at income $\langle x^0, p \rangle$ but which is better than x^*; that expenditure can always be thus 'traded' for satisfaction is assured by local non-satiation. In the proof of (ii) the bundle x^1 that is better than x^{**} is made a little smaller, lessening satisfaction in return for less cost, until a bundle is reached that is still as good as t but which costs less than x^{**}; such 'trading' of satisfaction for expenditure is guaranteed by the existence of locally cheaper points. However, if the expenditure on x^{**} at prices p is already least possible (that is, if x^{**} is p-minimal) then 'trading' in *that* direction cannot occur – one cannot go below least cost.

Of the five assumptions of the Theorem the only one whose meaning is not transparent and whose restriction is not 'reasonable' is (e), so that it comes as no surprise that the main thing wrong at the Arrow Corner is that (e) does not hold there. For further discussion of this Slater-like assumption and its role in general equilibrium theory, see consumption sets.

See Also

▶ Duality

Bibliography

Arrow, K.J. 1951. An extension of the basic theorems of classical welfare economics. In *Proceedings of the second Berkeley Symposium on Mathematical Statistics and Probability*, ed. J. Neyman. Berkeley: University of California Press.

Arrow, K.J., and G. Debreu. 1954. Existence of an equilibrium for a competitive economy. *Econometrica* 22: 265–290.

Bergstrom, T.C., R.P. Parks, and T. Rader. 1976. Preferences which have open graphs. *Journal of Mathematical Economics* 3: 265–268.

Deaton, A., and J. Muellbauer. 1980. *Economics and consumer behaviour*. Cambridge: Cambridge University Press.

Debreu, G. 1959. *Theory of value*. Cowles commission monograph no. 17. New York: Wiley.

McKenzie, L. 1957. Demand theory without a utility index. *Review of Economic Studies* 24: 185–189.

Newman, P. 1982. Mirrored pairs of optimization problems. *Economica* 49: 109–119.

Samuelson, P.A. 1947. *Foundations of economic analysis*. Cambridge, MA: Harvard University Press.

Cost of Production

John Eatwell

Adam Smith argued that competition would tend to establish the 'natural prices' of commodities produced, i.e. the prices at which 'the price of any commodity is neither more nor less than what is sufficient to pay the rent of land, the wages of labour, and the profit of stock employed. . .according to their natural rates' (Smith 1776, p. 65). In other words, the price of any produced commodity will, under the pressure of competition, be equal to its cost of production.

Ricardo was later to use the term 'cost of production' as a synonym for the 'value' of a commodity, where by value is meant, in the *Essay on Profits*, the difficulty or facility of production of the commodity, and in the *Principles*, the relative

quantity of labour necessary for the production of the commodity. The expression is used to represent the value of the commodity as manifest in terms of the standard of values used in the market. Ricardo dismissed the attempt by Malthus to draw a distinction between value and cost:

> Mr. Malthus appears to think that it is part of my doctrine, that the cost and value of a thing should be the same; – it is, if he means by cost, 'cost of production', including profits (1817, p. 47n).

Ricardo's identification of cost and value has led some authors to the belief that the classical theory of value is a 'cost of production' theory, which may be contrasted with the 'supply and demand' theory of neoclassical economics – an idea that was perhaps reinforced by Marshall's claim that classical theory was 'one-sided', was only one blade of the scissors:

> we might as reasonably dispute whether it is the upper or the under blade of a pair of scissors that cuts a piece of paper, as whether value is governed by utility or cost of production (1890, p. 548).

Walras, too, seems to suggest that 'cost of production' and demand are independent forces in the mutual determination of price.

In fact the condition that the cost of production of produced commodities is equal to their price is simply a definition of competitive equilibrium. The proposition that price is equal to cost is thus common to all theories of competitive value. The fact is not trivial, but is the outcome of competitive mobility. The meaningful question for any theory of value is what determines that cost and price.

The classical theory of value and distribution takes as data the size and composition of social output, the conditions of reproduction, and the real wage (Sraffa 1960; Garegnani 1984). In a model which contains only circulating capital these data may be formed into the following price equations:

$$Ap(1+r) = p \tag{1}$$

where A is the $n \times n$ input–output matrix, p the price vector and r the rate of profit. Wage goods are incorporated in the input coefficients. There is no presumption that A is invariant to changes in output.

Given that A is connected (all commodities are basic) then the equations may be solved for unique positive values of r and p. Although in (1) price is equal to cost of production, cost of production is not independent of price. Since A is connected the price of every commodity depends as much on its *own price* as on the price of other commodities (Sraffa 1960, ch. 2).

The neoclassical theory of value takes as its data the preferences of individuals, the technology, and the endowment of factor services. In equilibrium the price of each produced commodity will equal the sum of the rentals of the factor services used in its production, i.e. its cost of production:

$$Bw = p \tag{2}$$

where B is the $n \times m$ matrix of input coefficients of factor services, w is the vector of factor rentals, and p the vector of prices of commodities produced. In this case it might appear at first glance that cost of production is independent of price since prices on the left hand side of (2) do not appear on the right hand side. But the appearance is deceptive. The rentals which clear the markets for factor services are determined by the endowments of factor services and the demands for them, which are in turn functions of the prices of final products. Hence the prices of produced commodities are not 'determined' by their costs of production, but are determined simultaneously with their costs of production.

See Also

▶ Difficulty or Facility of Production

Bibliography

Garegnani, P. 1984. Value and distribution in the classical economists and Marx. *Oxford Economic Papers* 36(2): 291–325.

Marshall, A. 1890. *Principles of Economics*. London: Macmillan.

Ricardo, D. 1817. In *Principles of Political Economy and Taxation*, ed. P. Sraffa. Cambridge: Cambridge University Press , 1951.
Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations*. London: Methuen , 1961.
Sraffa, P. 1960. *Production of Commodities by Means of Commodities*. Cambridge: Cambridge University Press.

Cost–Benefit Analysis

David L. Weimer

Abstract

Cost–benefit analysis (CBA) is a collection of methods and rules for assessing the social costs and benefits of alternative public policies. It promotes efficiency by identifying the set of feasible projects that would yield the largest positive net benefits to society. The willingness of people to pay to gain or avoid policy impacts is the guiding principle for measuring benefits. Opportunity cost is the guiding principle for measuring costs. CBA requires that appropriate shadow prices be derived when policies have effects beyond those that can be taken into account as changes of prices or quantities in undistorted markets.

Keywords

Consumer surplus; Contingent valuation; Cost–benefit analysis; Distortions; Donor value; Equivalent variation; Hicks compensation; Hicks, John R.; Kaldor, N.; Marshallian demand curves; Opportunity cost; Option price; Present value; Pure time preference; Revealed preference; Shadow prices; Social choice; Social surplus; Substitutes and complements; Travel-cost method; Value of statistical life; Willingness to pay

JEL Classifications

D61

Public policies, such as infrastructure projects, social welfare programmes, tax laws and regulations, typically have diverse effects in the sense that people would be willing to pay something to obtain effects they view as desirable and would require compensation to accept voluntarily effects they view as undesirable. If, across all members of society, the total amount willing to be paid by those who enjoy desirable effects (benefits) exceeds the total amount needed to compensate those who suffer undesirable effects (costs), then adopting the policy would make it potentially possible to achieve a Pareto improvement on the status quo. If the benefits do not exceed the costs, then adopting the policy does not offer a potential Pareto improvement. How should such costs and benefits be determined? Cost–benefit analysis (CBA) is the collection of generally accepted methods and rules for assessing the social costs and benefits of alternative public policies.

The US Flood Control Act of 1936 appears to be the first call for CBA to be systematically used to inform public policy (Steiner 1974); it became embedded within modern welfare economics with articles by John R. Hicks (1939) and Nicholas Kaldor (1940) that set out the efficiency rationale for requiring policies to have positive net benefits. Two forces have contributed to the increased use of CBA since the 1960s. First, budget pressures and the desire to avoid inefficient regulations have led many governments to promote, or even require, the subjection of certain types of policies to CBA. Its use in the United States, particularly in the area of economic regulation, has been mandated by a series of Executive Orders (Hahn and Sunstein 2002). Her Majesty's Treasury in the United Kingdom publishes the *Green Book* to help public sector organizations apply CBA to ensure that 'public funds are spent on activities that provide the greatest benefits to society, and that they are spent in the most efficient way' (HM Treasury 2002: v). Second, economists have shown ingenuity in finding ways to value goods not traded in efficient markets, thereby expanding the range of policies to which CBA can be reasonably applied. For example, the travel-cost method provides a way to value recreational facilities that charge an administratively

determined entry fee (Clawson and Knetsch 1966); hedonic pricing models facilitate valuation of spatially varying local public goods (Smith and Huang 1995); and the development of the contingent valuation survey method, propelled by environmental damage assessment suits in US courts, permits the valuation of public goods, such as existence value, that lack readily observable behavioural traces needed for revealed preference estimation (David 1963; Bateman and Willis 2000).

CBA promotes efficiency by identifying the set of feasible projects that would yield the largest positive net benefits. Three conceptual criticisms can be made against this proposition. First, because those who suffer costs from a policy are almost never fully compensated, CBA in any particular application generally will not guarantee a Pareto improvement. The counter-argument is that, if CBA is consistently used to select policies offering the largest net benefits and there are no consistent losers, then it is likely that overall everyone will actually be made better off. Second, the CBA techniques for measuring net benefits cannot guarantee a coherent social ordering of policy alternatives. For example, it is possible to identify situations in which moving from one policy to another offers positive net benefits as does moving back to the original policy (Scitovsky 1941; Blackorby and Donaldson 1990). As no fair social choice rule can guarantee a transitive social ordering (Arrow 1963), this result is not surprising and is of minor consequence compared with the practical difficulties encountered in applying CBA. Third, and most important, only a few economists argue that public policies should be selected solely to promote the goal of efficiency. Other goals, such as equity and preservation of human dignity, are often legitimately viewed as relevant to policy choice, so that CBA is inappropriate as a decision rule. Nonetheless, as efficiency is almost always one of the relevant goals of public policy, CBA remains useful as a method for assessing efficiency in the context of a broader multi-goal analysis.

Social Perspective

CBA assesses social costs and benefits, which distinguishes it from the self- regarding calculus of individual economic actors. The meaning of 'social' in this context is twofold. First, it involves the definition of the relevant society; that is, it requires a determination of whose costs and benefits have standing (Whittington and MacRae 1986). Economists generally argue for national standing, recognizing that those in a particular country live under the same political contract, or constitution, and share a common economy with its own fiscal and monetary policy. In practice, however, sub-national governments often base their decisions only on their own costs and benefits and therefore demand CBA with standing restricted to those under their jurisdictions. Even when geographic standing is resolved, issues remain as to whether the costs and benefits of all residents – citizens, legal aliens, illegal aliens, those with legally proscribed preferences – should count (Zerbe 1998).

Second, it requires comprehensive assessment of the valued effects of policies on those with standing. The effects are commonly divided into the categories of active and passive use. Policies affect active use by changing the observable quantities of goods consumed, such as day care or fishing. Passive use includes all those effects that cannot be readily identified with observable changes in behaviour: existence value, or the willingness to pay for some good, such as wilderness, that one never expects to consume actively (Krutilla 1967); option value, or the willingness to pay for some good that one may wish to consume actively in the future (Weisbrod 1964); donor value, or the willingness to pay for redistributions of goods to others (Hochman and Rogers 1969). The absence of observable behaviour precludes valuation of passive use through the revealed preference methods most favoured by economists. Stated preference methods, such as contingent value surveys, are thus necessary for undertaking comprehensive assessments of policies with effects on passive use.

Social Benefits: Willingness to Pay

A common metric for policy effects is required if these effects are to be aggregated across individuals within the relevant society. If more than one policy alternative is to be compared with the status quo, then this metric must have ordinal properties. Further, if it is to be compared with the resource costs of implementing the policy, then it must be measured in the monetary unit of the society. Equivalent variation (*EV*) satisfies these conditions (McKenzie 1983). Consider the expenditure, or cost- utility, function C(U,P), where *C* is the amount of money needed to achieve utility *U* with price vector *P*. If U_1 is the person's utility under the price vector P_1 that would result from the policy change and P_0 is the price vector that would result under the status quo, then the equivalent variation of the policy change is given by

$$EV = C(U_1, P_0) - C(U_1, P_1)$$

the difference between the expenditure needed to achieve U_1 without the policy and with it. The EV is the amount of money that one would have to give to the person instead of implementing the policy so that the person is as well off as he or she would have been had the policy been implemented. A negative EV indicates that the person finds the net effects of the policy undesirable.

In its actual use, CBA almost always evaluates policy effects with willingness to pay, which differs conceptually from EV. Willingness to pay answers the question: how much money could be taken away from a person in conjunction with the policy so that he or she has the same utility with the policy as without it? Rather than corresponding to EV, which holds utility constant at a level with the policy, willingness to pay corresponds to compensating variation, which holds utility constant at the pre-policy level. Although compensating variation is more intuitively appealing, it does not provide a fully satisfactory money metric like EV.

The equivalent or compensating variation of a price change in a single market can be calculated as the change in social surplus as measured under the appropriate Hicksian, or utility-compensated, demand schedule. In practice, however, analysts typically work with econometrically derived demand curves that do not hold utility constant. Changes in consumer surplus measured with these Marshallian demand curves only approximate the compensating variation, with differences driven by income effects that can be large for either large income elasticities or large price changes. Some progress has been made to put bounds on the differences between the Marshallian and Hicksian measures (Willig 1976; Seade 1978), but these bounds are rarely applied in practice.

The interpretation of Marshallian consumer surplus as willingness to pay becomes even more complicated when policies have secondary effects in the markets of complements and substitutes of the goods primarily affected by policies. Although a general equilibrium model would be most appropriate for taking account of these secondary market effects, common practice is to approximate the combined effect of the primary and secondary markets by measuring surplus changes with the use of an estimated demand schedule for the primary market that does not hold the prices of substitutes and complements constant (Sugden and Williams 1978; Gramlich 1990; Boardman et al. 2006). In such cases, analysts need not account for price changes in undistorted secondary markets. Indeed, doing so would likely result in double counting of benefits.

Social Costs: Opportunity Costs

Public policies generally require the use of real resources to produce their effects. The guiding principle for monetizing the forgone value of these resources is opportunity cost: what is the value of the resources in their next-best use? That is, what is the value forgone by using the resources for the project? When factor markets are undistorted and the additional demand created by the project does not increase price, the opportunity cost of the resource just equals its market value,

which, if the resource is obtained by purchase, just equals the expenditure on the resource. When factor markets are undistorted but the additional demand induced by the policy drives up price, then the opportunity cost of the resource equals the sum of expenditure and the change in social surplus, the algebraic sum of the change in consumer surplus and the change in rents usually measured as change in producer surplus based on the short-run supply schedule (Mishan 1968). For example, if supply and demand curves are linear, then the opportunity cost equals the average of the pre- and post-purchase prices of the resource times the quantity purchased.

If markets are distorted, then even if price does not change the opportunity cost does not necessarily equal the expenditure required to secure supply. For example, a common factor-market distortion is involuntary unemployment resulting from minimum wages imposed by either law or custom. The expenditures needed to hire workers from a market with involuntary unemployment for a project clearly overestimate the opportunity cost of this labour. Nonetheless, the opportunity cost is almost certainly not zero, as sometimes argued by policy advocates, because the time of the workers hired by the project has an opportunity cost in terms of forgone leisure and household production.

Accommodating Uncertainty

CBA requires prediction of the effects of adopting a policy. Predictions are inherently uncertain. In addition to uncertainty about such parameters as price elasticities required for predictions of changes in social surplus, CBA often requires analysts to confront fundamental uncertainty about future states of the world. For example, preparing a vaccine to guard against a potential pandemic is costly but offers large benefits in the event that a pandemic actually materializes. CBA requires analysts to convert these uncertainties into risks by specifying representative states of the world and assigning probabilities to these states. Common practice is to model the policy choice as a decision analysis problem, or game against nature, and to choose the policy that maximizes the expected value of social surplus.

A more conceptually valid measure of the benefits of a project with certain costs in the face of risk about the future state of the world is option price (Graham 1981). Option price answers that question: what is the maximum certain payment that an individual would be willing to make to obtain the project? The sum of these certain payments for all those with standing can then be compared with the certain cost of implementing the policy. In general, however, option price does not equal the expected value of an individual's surplus over the possible states of the world; it differs from expected surplus by the option value of the policy for the individual. Although contingent valuation surveys seek to elicit individuals' option prices directly, more commonly analysts estimate benefits as expected surpluses, and consider option value as an excluded value. Some progress has been made in signing option value (Larson and Flacco 1992), but analysts rarely have enough information for confidently including it as a monetized correction to expected surplus.

Discounting for Time

Policies typically have effects that extend far into the future. Infrastructure projects in particular are usually characterized by large initial investments followed by beneficial use over years or even scores of years. CBA requires that costs and benefits accruing in the future be converted into their present value equivalents. On the assumption that future costs and benefits are predicted in real dollars, then a dollar of cost or benefit occurring t periods beyond the present equals in present value terms

$$1/(1+d)^t$$

where d is the real discount rate for the period length. In practice, discounting is usually done on an annual basis. As valid comparison of projects requires that they be assessed over the same time horizon, it is often necessary to convert present

values to equivalent perpetual streams of constant values through the use of an annuity factor.

The appropriate value for the real discount rate remains controversial. One approach is to set the discount rate equal to the marginal rate of pure time preference, the rate at which consumers are indifferent between exchanging current for future consumption. Another approach is to set the discount rate equal to the opportunity cost of capital, the marginal rate of return on private investment. In an ideal capital market these two rates would be equal. In the presence of transaction costs and taxes, however, these rates differ substantially. For example, an estimate of the marginal rate of pure time preference based on the after-tax real rate of return on US treasury bonds is 1.5%, while an estimate of the opportunity cost of capital based on the expected real yield on AAA corporate bonds is 4.5% (Moore et al. 2004).

If all costs and benefits correspond to changes in consumption, then the marginal rate of pure time preference is the appropriate discount rate. Instead, if all costs and benefits correspond to changes in private investment, then the marginal rate of return on private investment is the appropriate discount rate. However, most projects involve changes in both consumption and investment. The shadow price of capital approach involves expressing all costs and benefits in terms of consumption changes so that the marginal rate of pure time preference can be applied (Bradford 1975). In application, this means applying a shadow price to changes in private investment so that they are converted to the present values of their associated streams of consumption changes.

Shadow Prices

Much of the challenge of CBA lies in deriving appropriate shadow prices when policies have effects beyond those that can be taken into account as changes of prices or quantities in undistorted markets. In developing countries, for example, import and export controls and the presence of subsistence agriculture often distort virtually all prices, necessitating the determination of a complete set of shadow prices based on prices in international markets (Little and Mirlees 1974; Squire and van der Tak 1975; Dinwiddy and Teal 1996). Economic research provides a number of shadow price estimates that can be used in conducting CBA. Indeed, were these shadow prices not readily available, the plausible range of application of CBA would be much narrower.

One of the most commonly needed shadow prices is the value of a statistical life. That is, what is the willingness of a representative member of a population to pay for reductions in mortality risk? Economists have used a variety of methods to estimate the value of a statistical life, most commonly taking advantage of differences in risks and wages across occupations or the purchases of safety devices. The number of studies is sufficiently large that a number of meta-analyses have been conducted to develop estimates of the value of a statistical life for the United States in the range of roughly $4 million to $6 million in 2002 dollars (Miller 2000; Viscusi and Aldy 2003). Tied to any estimate of the value of a statistical life is the value of a life year. Health economists have developed a number of methods for estimating the quality of life in various health states, so that, in conjunction with the value of a life year, they can monetize a quality-adjusted life year (QALY) for use in CBAs of health care interventions (Dolan 2000). Estimates of shadow prices for injuries, noise, recreational activities, air pollutants, commuting time, and the marginal excess burden of taxation (for application to changes in government revenue) are also readily available (Boardman et al. 2006).

See Also

- ▶ Consumer Surplus
- ▶ Contingent Valuation
- ▶ Hedonic Prices
- ▶ Pareto Principle and Competing Principles
- ▶ Rent
- ▶ Social Discount Rate
- ▶ Value of Life
- ▶ Value of Time

Bibliography

Arrow, K. 1963. *Social choice and individual values*. 2nd ed. NewHaven: Yale University Press.

Bateman, I., and K. Willis, eds. 2000. *Valuing environmental preferences theory and practice of the contingent valuation method in the US EC and developing countries*. Oxford: Oxford University Press.

Blackorby, C., and D. Donaldson. 1990. A review article: The case against the use of the sum of compensating variation in cost–benefit analysis. *Canadian Journal of Economics* 23: 471–494.

Boardman, A., D. Greenberg, A. Vining, and D. Weimer. 2006. *Cost–benefit analysis: Concepts and practice*. 3rd ed. Upper Saddle River: Prentice Hall.

Bradford, D. 1975. Constraints on government investment opportunities and the choice of discount rate. *American Economic Review* 65: 887–899.

Clawson, M., and J. Knetsch. 1966. *Economics of outdoor recreation*. Baltimore: Johns Hopkins University Press.

David, R. 1963. Recreation planning as an economic problem. *Natural Resources Journal* 3: 239–249.

Dinwiddy, C., and F. Teal. 1996. *Principles of cost-benefit analysis for developing countries*. Cambridge: Cambridge University Press.

Dolan, P. 2000. The measurement of health-related quality of life for use in resource allocation in health care. In *Handbook of health economics 1B*, ed. A. Culyer and J. Newhouse. Amsterdam: Elsevier.

Graham, D. 1981. Cost–benefit analysis under uncertainty. *American Economic Review* 71: 715–725.

Gramlich, E. 1990. *A guide to benefit-cost analysis*. 2nd ed. Englewood Cliffs: Prentice-Hall.

Hahn, R., and C. Sunstein. 2002. A new executive order for improving federal regulation? Deeper and wider cost-benefit analysis. *University of Pennsylvania Law Review* 150: 1389–1552.

Hicks, J.R. 1939. The valuation of social income. *Economica* 7: 105–124.

HM Treasury. 2002. *The green book: Appraisal and evaluation in central government*. London: The Stationary Office.

Hochman, H., and J. Rogers. 1969. Pareto optimal redistribution. *American Economic Review* 59: 542–557.

Kaldor, N. 1940. Welfare propositions of economics and interpersonal comparisons of utility. *Economic Journal* 49: 549–552.

Krutilla, J. 1967. Conservation reconsidered. *American Economic Review* 57: 777–786.

Larson, D., and P. Flacco. 1992. Measuring option prices from market behavior. *Journal of Environmental Economics and Management* 22: 178–198.

Little, I., and J. Mirlees. 1974. *Project appraisal and planning for developing countries*. London: Heinemann Educational.

McKenzie, G. 1983. *Measuring economic welfare*. Cambridge: Cambridge University Press.

Miller, T. 2000. Variations between countries in the values of statistical life. *Transport Economics and Policy* 34: 169–188.

Mishan, E. 1968. What is producer's surplus? *American Economic Review* 58: 1269–1282.

Moore, M., A. Boardman, A. Vining, D. Weimer, and D. Greenberg. 2004. 'Just give me a number!' Practical values for the social discount rate. *Journal of Policy Analysis and Management* 23: 789–812.

Scitovsky, T. 1941. A note on welfare propositions in economics. *Review of Economic Studies* 41: 77–88.

Seade, J. 1978. Consumer's surplus and linearity of Engel curves. *Review of Economic Studies* 9: 77–88.

Smith, V., and J. Huang. 1995. Can markets value air quality? A meta analysis of hedonic property value models. *Journal of Political Economy* 103: 209–277.

Squire, L., and H. van der Tak. 1975. *Economic analysis of projects*. Baltimore: Johns Hopkins University Press.

Steiner, P. 1974. Public expenditure budgeting. In *The economics of public finance*, ed. A. Blinder et al. Washington, DC: The Brookings Institution.

Sugden, R., and A. Williams. 1978. *Principles of practical cost–benefit analysis*. Oxford: Oxford University Press.

Viscusi, W., and J. Aldy. 2003. The value of statistical life: A critical review of market estimates throughout the world. *Journal of Risk and Uncertainty* 27: 5–76.

Weisbrod, B. 1964. Collective consumption services of individual consumption goods. *Quarterly Journal of Economics* 78: 71–77.

Whittington, D., and D. MacRae. 1986. The issue of standing in cost–benefit analysis. *Journal of Policy Analysis and Management* 5: 665–682.

Willig, R. 1976. Consumer's surplus without apology. *American Economic Review* 66: 589–597.

Zerbe, R. Jr. 1998. Is cost–benefit analysis legal? Three rules. *Journal of Policy Analysis and Management* 17: 419–456.

Cost–Benefit Analysis: Philosophical Issues

Sven Ove Hansson

Abstract

Cost–benefit analysis (CBA) gives rise to a whole range of philosophical issues. The most discussed among these is the status of economic values that are assigned to assets conceived as incommensurable with money, such as a human life or the continued existence of an animal species. CBA also involves other

contentious assumptions, for instance that a disadvantage affecting one person can be fully compensated for by an advantage affecting some other person. Another controversial issue is whether a CBA should cover all aspects in a decision or rather leave out certain issues (such as justice) so that they can instead be treated separately.

Keywords

Aggregation; Commensurability; Comparability; Compensation; Contingent valuation; Cost–benefit analysis; Environmental economics; Ethics; Health economics; Incommensurability; Interpersonal comparison; Life value; Philosophy of economics; Risk–benefit analysis; Synopticism; Value of life; Welfare economics; Willingness to pay

JEL Classifications

D61

Cost–Benefit Analysis: Philosophical Issues

Cost–benefit analysis (CBA) is a collection of decision-aiding techniques that weigh advantages against disadvantages in numerical terms. In a typical CBA, multi-dimensional problems are reduced to one dimension, usually with monetary value as the common currency. Such a reduction raises several important philosophical issues (Hansson 2007; Sen 2000; Sunstein 2005).

Incommensurability

The most discussed among these issues concerns the status of the economic values that cost–benefit analysts assign to assets that do not have a market value. Many of these assets are conceived as invaluable, such as a human life or the continued existence of an animal species. Critics have claimed that CBA desecrates human life when it assigns a monetary value to the loss of human lives. Such criticism would probably have been less common if the nature of these values had been better explained. In particular, they are not prices. (No market – no price.) The assignment of a sum of money to the loss of a human life does not imply that someone can buy another person, or the right to kill her, at that price. What it implies is that society tends to pay (alternative: ought to pay) up to that sum to save a human life.

The incommensurability between life and money is only one of many incommensurabilities that are dealt with in CBA. There is no definite answer to the question how many cases of juvenile diabetes correspond to one death, or what amount of human suffering or death corresponds to the extinction of an antelope species. Since such comparisons are technically effected in a CBA by assigning monetary values, the problem of incommensurability appears to be a problem of monetisation. But even if money were removed from the analysis it would still be necessary to deal with comparisons between deaths, diseases and environmental damage. The fundamental problem is that for decision-making purposes we need to evaluate comparatively entities that we conceive as incomparable. Such 'impossible' comparisons are inherent in all major social decisions. CBA brings them to light.

Interpersonal Aggregation

In a CBA, all costs and all benefits are combined into one and the same balance. This means that a disadvantage affecting one person can be fully compensated for by an advantage affecting some other person. In other words, *interpersonal compensability* of advantages and disadvantages is assumed. (Interpersonal compensability should not be conflated with the related but distinct issue of interpersonal comparability. Even if a benefit to one person is greater than a harm to another person, it need not cancel out the harm.) The assumption of interpersonal compensability is one of several features that CBA analysis shares with utilitarian moral theory.

There is, at least theoretically, an alternative to this approach. Advantages and disadvantages can

be weighted against each other separately for each affected person, and a positive balance for each individual person can be required for a policy to be accepted. This is the approach that has dominated mainstream economics since the 1930s, when Lionel Robbins showed how economic analysis can dispense with interpersonal comparability. The approach that prevails in CBA is more akin to the collective, aggregating approach of the so-called old welfare economics. There is an obvious but surprisingly little discussed tension between standard CBA and Paretian welfare economics. The former, but not the latter, tends to sanction the sacrifice of individual interests for the sake of collective goals.

Many of the value assignments used in CBA are based on estimates or measurements of (hypothetical) willingness to pay. This applies for instance to values based on contingent valuation. All evaluation methods that are based on willingness to pay tend to give more influence to affluent people since they can pay more than others to have it their way. This can be corrected with income-based adjustments of the reported willingness to pay.

Exclusion of Aspects

All evaluations of the future effects of decisions tend, irrespective of methodology, to leave out or downplay effects that are difficult to predict. Furthermore, since CBA aims at numerical calculations, it tends to leave out aspects of future developments that can only be predicted in non-quantitative terms. This applies for instance to risks of cultural impoverishment, social isolation, and increased tensions between social strata. These limitations can lead to bias when alternatives with mostly quantifiable negative consequences are compared to alternatives whose major drawbacks are nonquantifiable. Furthermore, due to their aggregative structure, CBAs often leave out social justice and other distributional aspects from the analysis even when they are accessible to quantitative treatment.

Cost–benefit analysts have given two major answers to this criticism. One of these is that all such neglected factors could and should be included in the analysis. It is for instance not difficult to put a price on inequality and include it in the analysis, and the same applies to other aspects that are commonly left out. (However, such all-encompassing CBAs are much more seldom performed than they are referred to in defence of the CBA methodology.)

The other answer is that a CBA only covers some of the aspects of a decision. It should therefore not be treated as the last word in an issue, but has to be followed by reports and discussions that cover aspects not covered in the CBA. Some discussants consider it inappropriate to include distributive justice in the total calculations of a CBA, since such issues are better dealt with separately.

Transferability Across Contexts

In CBA, cost estimates are regularly transferred across contexts. This applies for instance to values of human life. A CBA that the U.S. Environmental Protection Agency performed for a new standard for arsenic in drinking water can be used as an example of this. The values of life used in this analysis were standard values derived from studies of how much male workers receive in compensation for risks of fatal accidents. However, as was noted by Heinzerling (2002), it was not necessary in this case to import life values from another context. It would have been possible to use life values derived from the very context of the CBA in question. There is a market for bottled, presumably non-toxic, water. Willingness to pay could have been derived from an analysis of prices on that market. Alternatively, consumers could have been asked how much they are prepared to pay for reduced levels of arsenic in drinking water, given realistic assumptions about the health effects of such a reduction.

The transfer across contexts that is illustrated in this example is an essential component in the methodology of CBA. If all values used in a CBAwere derived from the precise context of its

subject matter, then its usefulness for comparative purposes could be put in question. But even though transferability across contexts is an essential assumption in CBA, it is far from trivial to defend it from a philosophical point of view. Such a defence would have to show that our evaluations of a consequence should be the same irrespective of the context in which that consequence appears. For instance, a life lost in a workplace accident and a (statistical) life lost due to arsenic in drinking water should be assigned the same value.

In practice, we tend to pay much more to save a life in some contexts than in others. It is far from self-evident that all such differences lack sensible normative justification. It may for instance be justified to pay more to protect people from risks that they cannot avoid than to protect them against risks that they can avoid at small cost to themselves. For similar reasons, it may be justified to pay more to protect children than adults. Furthermore, some causes of death are considered particularly pernicious, and therefore worth more expensive countermeasures than other causes of death. We may for instance choose to pay more per life saved in a law enforcement programme that reduces the frequency of manslaughter than we would pay for most other life-saving activities.

Decisional Synopticism

The effects of a decision often depend heavily on other, parallel decisions. A CBA devoted to one of several interconnected decisions can be misleading due to the impact of the decisions that it does not cover. As one example of this, a CBA (or any other type of analysis) aimed at optimising the road traffic system may result in a suboptimal recommendation due to potentials of rail traffic that it does not take into account. Such effects of non-inclusion can in principle be remedied by framing decisions in large coordinated units that cover as many social areas as possible. The tendency to do so has been called 'super synopticism'. (Hornstein 1993, p. 387) However, such large-scale optimisation does not always work, largely for reasons similar to those that make centralised planning inefficient.

As one example of this, the willingness to pay for safety, as measured in the marginal cost for saving a life, differs widely between policy areas. Some cost–benefit analysts claim that all decisions on risk acceptance should be coordinated so that willingness to pay is equalised across policy areas. The implementation of such a unified price would require a high degree of coordination across policy areas. This is not easy to achieve since risk decisions are interwoven with other decisions in their respective policy areas. It may not always be feasible in practice to make risk decisions in a fully coordinated and centralised way while retaining a decentralised decision structure for other decisions.

This and several other issues connected with CBA will be much less problematic if a CBA is considered as one of several inputs into a decision than if it is presented as the last word which a rational decision-maker has to abide by.

See Also

- ▶ Contingent Valuation
- ▶ Cost–Benefit Analysis
- ▶ Ethics and Economics
- ▶ Philosophy and Economics
- ▶ Value of Life

Bibliography

Hansson, S.O. 2007. Philosophical problems in cost–benefit analysis. *Economics and Philosophy* 23: 163–183.

Heinzerling, L. 2002. Markets for arsenic. *Georgetown Law Journal* 90: 2311–2339.

Hornstein, D.T. 1993. Lessons from federal pesticide regulation on the paradigms and politics of environmental law reform. *Yale Journal on Regulation* 10: 369–446.

Sen, A. 2000. The discipline of cost–benefit analysis. *Journal of Legal Studies* 29: 931–952.

Sunstein, C.R. 2005. Cost–benefit analysis and the environment. *Ethics* 115: 351–385.

Cost-Push Inflation

George L. Perry

Abstract

The concept of cost-push inflation emerged after the Second World War to describe the price increases arising from labour unions pushing up wages despite excessive unemployment. With the oil price shocks of the 1970s, it was used to describe any important shift up in supply schedules at given levels of aggregate demand. Most central banks differentiate between supply shock effects and demand effects by distinguishing between overall inflation and core inflation, the latter omitting the direct contribution of shocks to oil and food prices, the two most important sources of supply shock large enough to register on broad inflation measures.

Keywords

Aggregate demand; Business cycle; Cost-push inflation; Excess demand; Expectations; Full employment; Incomes policies; Inflation; Labour supply; Learning; Market power; Monetary policy; Natural rate of unemployment; Organization of Petroleum Exporting Countries (OPEC); Phillips curve; Stabilization policy; Sticky prices; Supply shocks; Trade unions; Unemployment; Wage inflation; Wage rigidity

JEL Classifications

E3

The concept of cost-push inflation emerged in the period after the Second World War. The Keynesian model of that time emphasized that the economy could operate with inefficiently low utilization of its capital and labour resources, and that expanding demand would employ those resources. Once full employment was achieved, further expansion of demand would only pull up nominal wages and prices. In contrast to this demand-pull inflation, cost-push described the price increases that came from labour unions pushing up wages despite the existence of excessive unemployment. Since the 1970s, when oil prices rose by many times in two abrupt steps, the idea of cost push has been extended to describe price increases arising from any important shift up in supply schedules at given levels of aggregate demand.

The key distinction between price increases arising from monopoly power in wage settings or from any other supply shock and price increases arising from an increase in aggregate demand along an unchanged supply curve is important both for empirical modelling of inflation and for stabilization policy. By the 1960s, the short- run Phillips curve had emerged as a description of the relation between inflation and unemployment over the business cycle. It described an empirical regularity according to which wages rose gradually faster as unemployment declined, with the relation becoming steeper the lower the unemployment rate. Subsequent amendments to this model took explicit account of learning and expectations and of the interrelation between wages and prices. In the dominant model that emerged, inflation will accelerate (decelerate) indefinitely if the economy operates persistently below (above) a natural rate of unemployment. And in models that stress the importance of expectations, the anticipation of faster or slower price increases speeds up this process of acceleration or deceleration.

While inflation is responsive to aggregate demand in all these models, its responsiveness to supply shocks is more nuanced. In models that stress inflationary expectations, shocks that are widely perceived as one-time shifts up nominal supply curves will lead only to one-time shocks to the price level. In models with institutions that partially or fully index wages to prices, or models with adaptive expectations of inflation, such shocks will have larger and more protracted effects.

Empirically, the distinction between cost-push and excess-demand effects is not always easily drawn. The inflation identified with

unemployment below the natural rate or with the steep portion of the short-run Phillips curve is attributable to excess demand. The more modest variations in inflation that may occur as unemployment varies above the natural rate are not characterized so readily. A useful interpretation of these systematic cyclical tendencies is that they represent the normal operation of heterogeneous labour markets in response to cyclical variations in aggregate demand, with wages and prices in some sectors rising faster as their markets tighten while slack is still present in other sectors. On this interpretation, they neither signal that the economy is at a natural rate nor indicate the presence of exogenous cost-push effects on prices. However, these modest variations in inflation may also indicate cost-push effects in wage settings that interfere with achieving full employment, and at times past policymakers have interpreted them in this way and tried to suppress them.

The interdependence between prices and wages presents another difficulty in distinguishing endogenous from exogenous changes in wages. If labour supply depends on real wages – that is, wages relative to the average price level – then labour supply will not change if nominal wages change proportionally in response to disturbances to the cost of living. The narrowest concept of cost-push would, therefore, include only shifts up in labour supply schedules that raise wages relative both to their normal response to cyclical demand conditions and to their normal response to consumer prices.

Such complications obscure the possible presence of cost push from wages in typical circumstances. However, when the exercise of market power in wage setting is extreme, it becomes more apparent. In the United States, large wage increases in the early post-war years are examples of cost push. Coming after wartime controls, these did not raise the concerns that the abrupt acceleration of wages in many industrialized economies did in the late 1960s and early 1970s. For example, in Germany annual increases in hourly compensation jumped from 7.5 per cent in 1968 to 17.5 per cent in 1970, and in the United Kingdom the acceleration over the same period was from seven per cent to 15.5 per cent.

During the 1970s, supply shocks to important raw materials prices dominated world price developments in the decade, producing the second main type of cost- push inflation. These supply shocks included the historic increases in oil prices in 1973–4 and again in 1979, and the food price explosion of 1973. Although world aggregate demand was relatively strong in both 1973 and 1979, the magnitude of the price increases that resulted would not be expected, and is better seen as a consequence of major shifts in world supplies. A succession of poor crops provoked the food price rise, while the successful organization of the OPEC oil cartel, aided by a levelling off in United States oil production, caused the oil price explosion.

Coincident with the 1973–5 supply shocks, further large jumps in wage inflation occurred in several countries. In both the United Kingdom and Japan, annual increases in hourly compensation rose to more than 30 per cent from less than half that rate in 1972. Most other major industrial countries experienced similar, though less dramatic, accelerations in wages. Although these wage developments were doubtless fuelled by the effects of the supply shocks on consumer prices, the differences across countries indicate another round of wage push in many, even when one allows for a normal response to price changes. The rapid wage increases in turn further boosted consumer prices. The eventual changes in real wages, as well as the eventual increase in price levels, varied significantly among the industrial countries during the mid-1970s. In the United States, the speed-ups and slowdowns in wage increases and in prices were far less dramatic than in Europe and Japan. However, over the entire decade of the 1970s wages in the highly unionized sectors of the US economy outpaced economy-wide wages substantially, indicating a moderate but persistent wage push from important major industries.

While this post-war record shows that both wage push and supply shocks have at times been important in pushing up price levels, several difficulties remain with the idea of cost push as a distinct source of inflation, and some analysts reject the idea altogether. First of all, inflation

refers to an ongoing rate of increase in prices. A one-time rise in the average price level will translate into some rate of increase in prices over a period spanning the rise. Without quibbles over how long a time period is needed before a measurement qualifies as an 'inflation rate', the distinction between a one-time rise in the price level and an ongoing inflation rate is important. Second, inflation refers to the general price level, not to a subset of prices. A rise in oil prices is, first of all, a rise in the relative price of oil. If wages and prices were fully flexible and responded instantly to changes in the balance between demand and supply, then, in the presence of non-accommodating macroeconomic policies, cost-push shocks would indeed create only relative price changes; inflation, in the aggregate, would be impossible. Those who see monetary policy as able to control the overall price level, if not instantly at least over a relatively short period of time, see a cost push from some sectors as a relative price change that becomes a change in the overall price level only if accommodated by monetary policy. On this view, the accommodation rather than the cost-push causes the inflation.

However, such reasoning ignores the considerable downward rigidity in wages and stickiness in many prices as well as the interactions between prices and wages in modern economies, and thus loses the important role that cost-push shocks played in shaping economic performance in these inflationary periods. There are positive correlations among most prices and wages in the economy. In part these reflect common reactions to aggregate developments and in part they represent causal links among wages and prices throughout the economy.

When the links are strong, as they were in the inflationary periods of the late 1960s and 1970s, a cost-push supply shock will not only add directly to the average price level but will set in motion increases in other prices and wages strong enough to persist for some time, even in the face of slowing demand and increasing underutilization of resources. Consequently, an attempt by monetary policy to hold the overall price level unchanged in the face of such a cost-push shock will result mainly in reducing output and employment. Only gradually will the upward movement of prices originating from a supply shock yield to restrictive monetary policies. On the other hand, because the initial shock induces positively correlated responses in wages and other prices, an accommodative policy that aims to maintain output and employment in the face of the shock will result in a rise in the overall price level that is substantially greater than the direct effect of the shock itself. The question confronting stabilization policy is thus how much to accommodate. And the best answer will differ with different institutions and at different times.

The idea that cost-push inflation originating in excessive union wage demands would interfere with the attainment of full employment prompted attempts in several countries to design incomes policies as part of the stabilization policy arsenal. The idea was that demand management by government would aim at keeping the economy around full employment, while understandings among government, labour and business would aim at heading off wage-push inflation that might otherwise arise before full employment was achieved. There was some evidence of success from incomes policies, known as wage-price guideposts in the United States, in the mid-1960s (Perry 1967). But whatever chance such policies may have had in the longer run in a relatively benign environment, they were overwhelmed once economies were driven into the excess demand region during the Vietnam War, and the oil and food supply shocks of the early 1970s sharply raised average price levels everywhere. There has been little interest in incomes policies since that time.

By the 1990s, conventional stabilization policies had achieved low inflation rates throughout the industrial world, and the power of unions to originate more inflationary wage increases was very sharply reduced in almost all countries. Both these developments have lessened the problems of stabilization policy. There is evidence that the low-inflation environment has sharply reduced the links that formerly caused price shocks to spark a wage-price inflationary spiral, as they did in the 1970s (Brainard and Perry 2000). Wages did not accelerate in response to the world oil price shocks

of the mid-2000s, and monetary policymakers were able to focus largely on the core inflation rate – the aggregate inflation rate excluding food and energy prices – in setting policy. At least for now, inflation originating from cost push poses a much smaller risk for stabilization policies today than it has at times in the past.

See Also

- ▶ Demand-Pull Inflation
- ▶ Inflation
- ▶ Inflation Expectations

Bibliography

Akerlof, G., W. Dickens, and G. Perry. 1996. The macroeconomics of low inflation. *Brookings Papers on Economic Activity* 1996(1): 1–59.

Brainard, W., and G. Perry. 2000. Making policy in a changing world. In *Economic events, ideas, and policies*, ed. J. Tobin. Washington, DC: Brookings Institution.

Okun, A. 1981. *Prices and quantities: A macroeconomic analysis*. Washington, DC: Brookings Institution.

Okun, A., and G. Perry. 1978. *Curing chronic inflation*. Washington, DC: Brookings Institution. Also available as G. Perry, Slowing the wage-price spiral: the macroeconomic view. *Brookings Papers on Economic Activity* 1978(2), 259–99.

Perry, G. 1967. Wages and the guideposts. *American Economic Review* 57: 897–904.

Schultze, C. 1985. Microeconomic efficiency and nominal wage stickiness. *American Economic Review* 75: 1–15.

Tobin, J. 1972. Inflation and unemployment. *American Economic Review* 62: 1–18.

Counterfactuals

Donald N. McCloskey

Counterfactuals are what ifs, thought experiments, *Gedankenexperimenten,* alternatives to actual history; they imagine what would have happened to an economy if, contrary to fact, some present condition were changed; in the philosophical literature therefore they are known also as 'contrary-to-fact conditionals'.

The notion has been used most self-consciously in historical economics. For example: 'If railroads had not been invented the national income of the United States in 1890 would have been at most 5% lower.' Counterfactuals are implied, however, in many other parts of economics, such as macroeconomics: 'If a monetary rule with a small growth rate of M_1 were adopted then the rate of inflation would fall.' Or industrial organization: 'If the instant camera industry had 100 suppliers it would be competitive.'

The philosophical problem that counterfactuals raise, and part of the reason they have attracted the attention of modern philosophers, can be seen in the last example. We wish to contrast the present monopoly of instant cameras with (nearly) perfect competition. Perhaps we wish to do so in order to measure the welfare cost of the monopoly and to advise a judge. Now of course if somehow the instant camera industry were to have 100 sellers then each seller would be small relative to the whole demand or supply. Speaking mechanically, the usual formulas for elasticities imply that the elasticity of individual demand facing any one of them would be large, roughly 100 times the elasticity of total supply plus 100 times the elasticity of total demand. Such calculations are the heart of applied economics: If the cigarette tax were lowered what would be the new relative price of cigarettes? If the money supply were increased what would happen to the price level? If foreign doctors could practise freely in the United States what would happen to the cost of American medical care?

Such questions involve looking into a world having, say, an instant camera industry with 100 sellers rather than one. It would not be our world, which saw the miraculous birth of Polaroid, the struggle with Kodak, and the final triumph of patent over antitrust law. So much is clear. But how then is the counterfactual world to be imagined? A world in which the conditions of technology, personality, and law resulted in 100 Edwin Lands and 100 miniature Polaroid companies would be a different one – there's the condition contrary to fact.

The problems which can afflict counterfactuals are two: vagueness and absurdity. The vagueness arises when the model has not been fully specified. The world could arrive at 100 companies in many different ways, each with different implications for the original question about welfare. One can imagine getting 100 Polaroid companies, for example, by fragmenting edict now, well after the invention, in the style of the American Telephone and Telegraph case. Whatever the advantages, there might be inefficiencies in this. It would certainly change the future patent law. The change in law would in turn change things for good or ill elsewhere in the economy. A world in which patents are granted and then prematurely abrogated differs from the present world. Alternatively one might imagine subsidies in the 1940s that would have resulted originally in 100 alternative technologies of instant cameras (though actually only two were invented). This counterfactual likewise would have its costs, though different ones, changing for example the expectations of inventors about subsidies. A counterfactual requires a model broad enough to do the job.

Vagueness is solved by explicitness. The conditions required for various counterfactuals are made explicit, and being explicit can be tested for plausibility. Historical economists have been making counterfactuals explicit since the 1960s, using them to explore the causes of the American revolution and the consequences of American slavery (the counterfactual work is well surveyed by McClelland 1975).

In the most famous use of counterfactuals Robert W. Fogel (1964) calculated what the transport system of the United States in 1890 would have looked like without railroads. He argued that evaluating the 'indispensability' of the railroads entailed calculating what American life would have been like without them. Some historians were reluctant to talk about such a counterfactual, saying that it was "as if" history, quasi-history, fictitious history – that is not really history at all . . ., a figment' (Redlich 1968, in Andreano (ed.), pp. 95f). But economists find the notion natural, and philosophers accept it as routine. Indeed, the philosophers point out that the following are nearly equivalent (Goodman 1965, p. 44):

Scientific Law: All inflations arise from money growth.

Causal Assertion: Money growth alone causes inflation.

Factual Conditional: Since inflation has changed, money growth has changed.

Dispositional Statement: Inflation is controllable with money growth.

Parallel Worlds: In a world identical (or sufficiently similar) to ours except that money growth differed, inflation would be different.

Counterfactual: If money growth were to be held at zero, inflation would be zero.

The philosophy of counterfactuals revolves around the translation of one of these into another. Historians, not realizing that one is translatable into the other, flee the counterfactual in terror and cling to the causal statement. Yet economists have on this score no cause for smugness, since they have parallel philosophical fears. Economists flee the causal statement as historians flee the counterfactual, and believe as historians do that the thing itself can be avoided by suppressing its name.

Fogel's calculations stirred great controversy, but were robust (Fogel 1979). Since he was interested in long-term economic growth he did not imagine a sudden closure of the railroads in 1890: that clearly would have resulted in a very large drop in national income. Mental experiments like this commonly lie behind claims that railroads (or airlines or postal services or garbage collection) are 'essential'. Fogel imagined instead what the American economy would have looked like without access to railroads from the beginning, forced from the 1830s onward to rely on substitutes.

Such an economy would have invested more in canals and roads (Fogel introduced some of these into his counterfactual world, using contemporary engineering studies proposing them). It would have been an economy closer to waterways, with a bigger St. Louis and a smaller Denver. It would doubtless have invented more improvements in road transport, arriving at internal combustion a little earlier than the world we know.

Fogel could not specify every feature of the 'true' counterfactual world. But he suspected anyway that the true counterfactual would give a national income only a little below the actual. To test the suspicion, therefore, he biased the case against himself, choosing a 'practical' counterfactual world in which income would be if anything lower than in the true counterfactual: he did not introduce the internal combustion engine before its time; and he did not shift the location of the population to accommodate the non-railroad transportation. He forced his practical counterfactual to carry supplies by river, canal and horse cart (not by the motor trucks that might have been) to a Denver no smaller than it actually became at the height of the railroad age. The result was a calculable upper bound on the true impact on national income: since the 'true' counterfactual would have economized relative to the clumsy 'practical' counterfactual, a use of the practical counterfactual biases the case against a large impact. Fogel reckoned that the impact was at most five per cent of 1890 income, a couple of years of economic growth.

He was merely applying in a bold way the usual methods of economics. The usual method is to imagine an explicit economic model, *M*, with parameters, *P*, and initial conditions (or exogeneous variables), *I*, and results by way of endogenous variables, *R*. The counterfactual varies some element of the setup, the simplest being a variation in *I* – where *I* might be a tax rate in a model of cigarette consumption or the number of firms in a naive model of instant camera pricing – and examines the results. Fogel removed from the initial conditions one of the technologies of transportation. In similar fashion a 500-equation model of the American economy permits experimentation in counterfactual worlds: What would happen if the price of oil fell? What would be the effect of a tax change? (The main empirical attack on Fogel's finding, indeed, was an highly explicit general equilibrium model of the Midwest and East (Williamson 1974).)

Counterfactuals are one of the two main ways that economists at present explore the world (the third, controlled experiment, is still not common). The first is regression, or the comparative method, asking how *in fact* results have varied with initial or exogenous conditions. The second is the counterfactual, or simulation, asking how the results *would* vary. The regression infers parameters *P* from data on initial conditions *I* and results *R* and from arguments about the model, *M*; the counterfactual simulation infers *R* from data on *P* and from arguments about *M* and *I*.

But in solving the vagueness of counterfactuals by positing explicit models the economist runs against the other philosophical problem of counterfactuals: absurdity. Consider again the counterfactual of a 100-firm industry selling instant cameras. The problem is that the initial conditions that would lead to such an industry may themselves be absurd. Indeed, they may violate the very model used. The counterfactual assertion 'If the instant camera industry were perfectly competitive then price would be lower than it is now' takes on the character of the proverbial line 'If my grandmother had wheels she'd be a tram.' The model may be true (wheeled grandmothers may indeed be trams) but the counterfactual may be impossible – that is, a contradiction of the model itself or of some other, wider model felt to be persuasive.

It is possible to argue on these grounds that *all* counterfactuals are absurd. One might argue, as did Leibniz, that a world that did not invent the railroad would strictly speaking have to be a world different from ours right back to the big bang. Such a world might be one in which the seas were boiling hot or pigs had wings, with different transportation problems. The theory being violated by the counterfactual is the theory that the world hangs tightly together. As J.S. Mill remarked in attacking counterfactual comparison of free trade and protection, 'Two nations which agreed in everything except their commercial policy would agree also in that' (1872, p. 575).

A less intense scepticism on the matter has figured widely in economics. The theory of games, for example, can be viewed as an inquiry into counterfactuals, which sometimes violate wider theories (Selten and Leopold 1982); the usual criticisms of the Cournot solution made by students of industrial organization involve the same point. Most notably, the Lucas Critique of

econometric policy evaluation (Lucas 1976) can be restated as a criticism of the usual counterfactual. The usual counterfactual imagines the effects of a change in the initial conditions *I* on a model *M* with given parameters *P*, fitted under the old regime. A new monetary policy would change the regime under which people believed they operated, changing *P* and *M* as much as *I*. Some broader model of how people adjust to regime changes is necessary to decide which would change: a new policy believed to be temporary would have very different effects from one believed to signal a revolution in government. The usual counterfactual violates the broader model, by supposing that people do not anticipate changes of regime or understand them when they occur. A broader model of rational expectations shows the counterfactual to be absurd.

John Elster, in a penetrating discussion of the role of counterfactuals in the economic sciences, posed the Basic Paradox of Counterfactuals: the less vague the theory, the more likely is a counterfactual using the theory to encounter absurdity. If Fogel had developed a theory of invention to draw a less vague picture of road transport without railroads he would have faced the problem that the very theory would predict the existence of railroads. After all, railroads were actually invented and therefore should be predicted by a sound theory of innovation. Elster wrote, 'If he attempted to strengthen his conclusion … he would be sawing off the branch he is sitting on. In this kind of exercise it is often the case that more is less and that ignorance is strength' (1978, p. 206). The counterfactual must be 'capable of insertion into the real past'.

The Basic Paradox illuminates the discussion in economics about simplicity of models. A simpler model is harder to believe in its simulation because it is not so rich; but because of its lack of richness it is more likely to be insertable into the real past. A 500-equation model of the economy will more tightly constrain the past from which it comes than will a 10-equation model. Model selection has its own type I and type II errors.

Many of the meta-criticisms of economics, then, reduce to remarks about a counterfactual. This is scarcely odd, since counterfactuals are equivalent to causal statements and the point of economics is to make causal statements. The philosophical literature on counterfactuals is illuminating, though large, technical, and mainly inconclusive (Lewis 1973; Goodman 1965). It comes to a position more sophisticated than mere scepticism. Counterfactuals are a way economists speak, and philosophers wish usually to assist the speaking, not end it. Self-aware or not, economists will go on speaking counterfactually about non-cooperative games, macroeconomic policy, and the retrospective welfare calculations of historical economics. The task of a philosophy of the economic counterfactual would be to understand the practice, not to change it.

See Also

- ▶ Cliometrics
- ▶ Models and Theory
- ▶ Philosophy and Economics

References

Elster, J. 1978. *Logic and society: Contradictions and possible worlds*. New York: Wiley.

Fogel, R.W. 1964. *Railroads and American Economic Growth: Essays in econometric history.* Baltimore: Johns Hopkins Press.

Fogel, R.W. 1979. Notes on the social saving controversy. *Journal of Economic History* 39(1): 1–54.

Goodman, N. 1965. *Fact, fiction and forecast*, 2nd ed. Indianapolis: Bobbs-Merril.

Lewis, D.K. 1973. *Counterfactuals*. Cambridge: Harvard University Press.

Lucas, R.E. 1976. Econometric policy evaluation: A critique. *Journal of Monetary Economics, Supplementary Series* 1: 19–46.

McClelland, P.D. 1975. *Causal explanation and model building in history, economics, and the new economic history.* Ithaca: Cornell University Press.

Mill, J.S. 1872. *A system of logic*, 8th ed. London: Longmans; reprinted, 1956.

Redlich, F. 1968. Potentialities and pitfalls in economic history. *Explorations in Entrepreneurial History* II, 6(1), 93–108. Reprinted in *The new economic history: Recent papers on methodology,* ed. R.L. Andreano. New York: Wiley, 1970.

Selten, R., and U. Leopold. 1982. Subjective conditionals in decision and game theory. In *Studies in contemporary economics*. Berlin: Springer.

Williamson, J.G. 1974. *Late Nineteenth-Century American Development: A general equilibrium history*. Cambridge: Cambridge University Press.

Countertrade

Dalia Marin and Monika Schnitzer

Abstract

International countertrade – tying an import to an export – emerged in the 1980s in response to the international debt crisis. Barter – the exchange of goods without using money – re-emerged in transition economies in the 1990s, in response to a domestic debt crisis. Both phenomena can be explained as institutional responses to contractual problems arising in imperfect capital and goods markets. Countertrade introduces a deal-specific collateral that improves the creditworthiness of countries and firms, and facilitates technology transfer to developing countries. Barter helps to overcome the lack of trust problem in the former Soviet Union.

Keywords

Asymmetric information; Barter; Buyback; Collateral, deal-specific; Commitment; Contract enforcement; Counterpurchase; Countertrade; Creditworthiness; Credit constraint; Cross-subsidy; Foreign direct investment; Foreign exchange shortage; Incentive contracts; International debt crisis; Liquidity constraints; North–South economic relations; Planning; Price discrimination; Reputation; Social networks; Social norms; Soft budget constraint; Transfer of technology; Trust, lack of; Virtual economy

JEL Classifications

F

Countertrade is a commercial transaction in which a seller, typically from an industrialized country, supplies goods, services or technology to a buyer in a developing country or a formerly planned economy, and in which, in return, the seller purchases from the buyer an agreed amount of goods, services or technology. A distinctive feature of countertrade is the existence of a link between the two transactions, the original import in the developing country and the subsequent export.

Countertrade takes a variety of forms. The three most commonly distinguished are 'barter', 'counterpurchase' and 'buyback'. Barter in the strict sense of the word refers to an import that is paid entirely or partly with an export from the importing country without using foreign exchange. Counterpurchase refers to a transaction in which the import is paid with foreign exchange, but the industrialized country commits to buy export goods from the developing country in return. Buyback is a transaction in which the seller supplies a production facility and the parties agree that the supplier of the facility will buy goods produced with that production facility. All three forms of countertrade are frequently observed in international trade.

Under central planning, countertrade was especially observed in international trade among countries belonging to the Council for Mutual Economic Assistance (Comecon, an economic organization of communist states) as well as in East–West trade. In particular in the 1980s, in the aftermath of the international debt crisis, countertrade became prevalent in international trade with developing countries and Eastern Europe. Before 1989 countertrade accounted for up to 40 per cent of total trade between East and West. After 1989, with the domestic debt crisis in transition countries, barter became dominant in domestic trade in these countries. While countertrade continues to be significant in North–South trade, reliable estimates are not available.

Explanations for Countertrade

One of the most frequently cited explanations of countertrade is that it allows countries to

overcome a shortage of hard currency. The observation that countertrading countries are highly indebted is taken as evidence that these countries face a shortage of foreign exchange and that their low creditworthiness makes it impossible to finance imports with a simple loan from an international bank (for example, OECD 1981, 1985). This interpretation is not fully plausible because countertrade uses export goods which otherwise could have been used to generate foreign exchange to pay for future imports. Furthermore, if the foreign-exchange shortage were the main explanation of countertrade we would expect barter to be the prevalent form of countertrade since only barter does in fact avoid the use of hard currency. However, barter accounts for only a small portion of total international countertrade (Marin and Schnitzer 2002a). Mirus and Yeung (1987) find that countertrade in the form of simple barter or counterpurchase does not improve a country's foreign exchange position unless it improves economic efficiency in the sense that it leads to an increase in national income.

Empirical evidence points to another explanation, starting from the observation that in international trade contract enforcement is problematic and hence conventional contracts cannot be relied on as the main mechanism to sustain economic exchange. International countertrade (as well as domestic barter, as pointed out below) can be explained as an institutional response to such contractual problems arising in imperfect capital and goods markets. Difficulties in contract enforcement are an important impediment to international transactions in the world economy. In international trade, national sovereignty interferes with contract enforcement because national borders demarcate national jurisdictions. Such demarcations segment markets and impose severe transaction costs on exchanges across national jurisdictions. The hazards involved in international transactions are often disregarded, but they make headlines each time a sovereign debtor threatens to stop servicing its debt, as it happened in the international debt crisis in the 1980s or in the Russian financial crisis in 1998.

If contract enforcement is weak, problems may arise on both ends of a business transaction: the seller may fail to deliver the goods, and the buyer may fail to pay for them. If buyers have no cash to pay, and thus face liquidity constraints at the time of delivery, the business transaction can take place only if the seller can trust the buyer to pay in due course. On the other hand, the buyer is willing to engage in a business transaction only if she can trust the seller to deliver the right goods. Both problems are prevalent in international trade. Enforcing the payment of goods can pose serious problems. In the aftermath of the debt crisis, highly indebted countries were liquidity constrained and could not finance necessary imports. Given their level of indebtedness, debt repayment could not be relied on. The debtor country could create more liquidity by not repaying its debt rather than by receiving a new loan. There are also important problems arising on the seller's side. In international trade, the most conspicuous example is the technology transfer problem. It is often reported that explicit contracts cannot be relied on to make sure that developing countries receive the advanced technology promised (Parsons 1987; Kogut 1986). These countries often complain that firms from industrialized countries sell inferior technology to them, technology that is outdated and cannot be sold on Western markets.

Solving the Creditworthiness Problem

Countertrade can be interpreted as the institutional response to the lack of creditworthiness of countries and firms. Countertrade introduces a deal-specific collateral for the credit granted for the original import. This collateral protects the interests of the creditor for one particular business transaction and thus mitigates the contractual hazards associated with indebtedness that would otherwise prevent the transaction from taking place.

The argument that payments in kind may have advantages over payments in cash contradicts the conventional wisdom in the theory of money. The common view is that barter is inefficient because it does not overcome the 'double coincidence of wants problem' (where each trading partner wants to buy exactly what the other partner wants to sell

and vice versa) as money does. A seller may need to accept goods for which she has no use herself. The point, however, is that goods have superior credit enforcement properties to those of money. Money is an anonymous medium of exchange. This anonymity can prove disadvantageous in trade with countries which lack creditworthiness, since the debtor in the developing country or eastern Europe can use it for purposes other than repaying debt. Goods, by contrast, can more easily be earmarked as the property of the creditor and can thus serve as collateral. However, payment in goods is problematic if it is difficult to judge the quality of the goods offered as means of payment. Thus, it is important to choose goods that are very liquid and hardly anonymous, making it both easy to determine their value and easy to earmark. Goods can be ranked with respect to their liquidity and anonymity properties, providing an explanation for the export pattern of countertrade and barter (Marin and Schnitzer 2002b).

Solving the Technology Transfer Problem

Buyback contracts have been interpreted as incentive contracts that ensure the transfer of desirable quality technology and post-installation service performance if standard forms of internalization, like joint ventures or foreign direct investment, are not possible due to political and ownership constraints (Hennart 1989; Chan and Hoy 1991; Mirus and Yeung 1993). But for the argument to work, it is essential that there be a technological relation between the two goods to be traded. However, buyback accounts for a surprisingly small fraction of all countertrade transactions. Thus, even though this explanation is theoretically appealing, it cannot explain the great majority of technology imports, which take the form of counterpurchase.

Interestingly, the technology transfer problem can be solved with a simple counterpurchase transaction as well, despite the lack of a technological link (Marin and Schnitzer 1995). Although the lack of liquidity makes it difficult to finance imports, it is this very lack of liquidity that can actually help when it comes to dealing with problems on the supplier's side. The idea is that the export from the developing country serves as a hostage that deters cheating on technology quality and defaulting on the payment of the original import from the industrialized country. For this mechanism to work, the export has to be profitable to both the industrialized country firm and developing country, and the contract is so designed that the export becomes sufficiently less profitable for either party that does not fulfil its obligations in the original import, be it technology transfer or payment of the import. The technology seller offers high-quality technology because otherwise she loses her collateral for the credit as the developing country firm would lack the revenues that are generated with the technology and that are necessary to produce the export goods. This contractual arrangement makes the technology supplier internalize the externality her technology imposes on the developing country. The developing country party will deliver the export goods because the terms of the contract are designed such that this is more profitable than selling them otherwise. So although the import and the export are not technologically related, the countertrade contract establishes a financial link that improves the incentives of the parties involved. Thus, countertrade is a first-best substitute for foreign direct investment when these countries are reluctant to give access to foreign ownership in their markets. This goes to prove that, in an imperfect world in which contract enforcement is weak, as in developing countries or imperfect capital markets, something that seems to be worse – that contractual problems arise on both sides of the business transaction rather than on only one side – can improve contract enforcement. In international trade, the liquidity constraint helps to solve the technology transfer problem.

Other Explanations

Some other possible explanations of countertrade are that developing countries use

countertrade transactions to promote the export of 'new' goods – goods they have not previously exported to industrialized countries – in order to gain access to new markets and to diversify their exports (OECD 1981, 1985). The empirical evidence gives some support for the view that countertrade has helped to stimulate and diversify exports. Other studies confirm that the goods exported by developing countries through countertrade arrangements are often goods for which export markets have yet to be established. Readily marketable products, like raw materials, are usually not available for countertrade. It can also be observed that a country removes goods from the countertrade shopping list once it has gained some experience with exporting these particular goods (Banks 1983). Furthermore, it has been argued that countertrade corrects distortions in non-competitive markets (Caves 1974). Using barter may allow competing more aggressively without openly violating collusive agreements. It may also allow more effective price discrimination. There is indeed some evidence that barter is used as a vehicle to change the terms of trade to allow price discrimination by Western monopolists (Caves and Marin 1992). Mandated countertrade has also been discussed as a policy response to contracting failures arising from asymmetric information about goods valuations (Ellingson and Stole 1996).

Barter Trade in Transition Economies

Barter trade has received renewed attention in the 1990s, when it became a dominant phenomenon in domestic trade in a number of transition countries, most notably in the successor states of the former Soviet Union. After 1989, domestic barter in Russia increased manifold after macroeconomic stabilization in 1994, from five per cent of GDP to 60 per cent in 1998. In Ukraine, the share of barter in industrial sales is estimated to have been more than 50 per cent in 1997. Only since the financial crisis in August 1998 have barter and the use of other money surrogates started to decline again.

A number of different explanations have been put forward for this phenomenon. Some experts have viewed it as a tax-avoidance mechanism because it allows a distortion of the true value of profits, and thus reduces tax liabilities. Furthermore, since the banking sector acts as a tax collection agency that transfers firms' cash income in bank accounts to the state to pay for outstanding tax arrears, barter allows tax avoidance because it avoids payments in cash. While there may be some truth in this kind of argument, few firms report tax advantages as a major reason for using barter (Marin and Schnitzer 2002a).

A more popular explanation refers to soft budget constraints and the lack of market discipline. The absence of hard budget constraints, so the argument goes, leads managers and workers to avoid the costs arising from restructuring by maintaining production in inefficient activities. Barter would allow concealing the true market value of output. But the empirical evidence suggests that barter is not a phenomenon of state-owned enterprises. Newly established private firms display an exposure to barter that is similar to or greater than that of state-owned firms or cooperatives (Marin and Schnitzer 2002a).

The 'virtual economy' argument of Gaddy and Ickes (1998) has been one of the most influential explanations of barter in Russia. The virtual economy argument claims that barter helps to create the image that the manufacturing sector in Russia is producing value while in fact it is not. This argument rests on the assumption that the manufacturing sector is value-subtracting, and most participants in the economy have an interest to pretend that it is not. Barter allows the parties to keep up this illusion by allowing the manufacturing sector to sell its output at a higher price than its market value and the value-adding natural resource sector (Gazprom) to accept this high price because of a lack of other sources. This way the manufacturing sector survives by drawing resources from the natural resource sector. According to the argument, keeping up the illusion of a value-adding manufacturing sector is highly costly for the Russian economy at large because this cross-subsidizing from the value-adding natural resource sector to the value-

subtracting manufacturing sector prevents the manufacturing sector from moving into valuable activity.

This argument appeals to experts of central planning and policy observers in transition economies, because the practice of cross-subsidizing across different activities in the economy was a widespread feature of central planning. But it raises a number of questions. If the natural resource sector is producing valuable output, why does the sector not have other opportunities than to subsidize the manufacturing sector? In fact, the natural resource sector is supposed to have significant bargaining power in the interaction with other sectors when it is producing goods which the market values highly. Why then does the sector end up subsidizing the rest of the economy? And in fact, evidence from barter transactions in the Ukraine suggests that, in contrast to the assertions of the virtual economy proponents, the electricity and gas industries in the natural resource sector gained from barter transactions, instead of losing (Marin 2002).

A more plausible explanation refers to the similarities between barter in international trade and barter in transition economies, and links the surge of domestic barter to a 'lack of trust' problem (Marin and Schnitzer 2005). In transition countries, poorly developed legal and financial institutions made contract enforcement unreliable and imposed severe transaction costs on any economic activity. These costs became prohibitively large in times of historic change and revolution. Unstable business partner relationships and rapidly changing social norms limited the extent to which economic exchanges could be sustained by reputation, by repeated interactions or by embedding them in social networks. This led to a lack of trust, meaning that reliable input supplies on the one hand and credit enforcement on the other hand were difficult to sustain, resulting in economic disorganization and a tremendous output fall. In such an environment, barter can be used as a commitment device to overcome the problems of unreliable input supplies and credit enforcement, by linking transactions and specifying terms of trade that give the right incentives to adhere to the terms of the barter contract.

See Also

- ▶ Barter
- ▶ International Trade Theory
- ▶ Planning
- ▶ Third World Debt
- ▶ Transfer of Technology

Bibliography

Banks, G. 1983. The economics and politics of countertrade. *World Economy* 6: 159–182.

Caves, R. 1974. The economics of reciprocity: Theory and evidence on bilateral trading arrangements. In *International trade and finance: Essays in honour of Jan Tinbergen*, ed. W. Sellekaerts. London: Macmillan.

Caves, R., and D. Marin. 1992. Countertrade transactions: Theory and evidence. *Economic Journal* 102: 1171–1183.

Chan, R., and M. Hoy. 1991. East–West joint ventures and buyback contracts. *Journal of International Economics* 30: 331–343.

Ellingson, T., and L.A. Stole. 1996. Mandated countertrade as a strategic commitment. *Journal of International Economics* 40: 67–84.

Gaddy, C.G., and B.W. Ickes. 1998. Russia's virtual economy. *Foreign Affairs* 77: 53–67.

Hennart, J.-F. 1989. The transaction-cost rationale for countertrade. *Journal of Law, Economics and Organization* 5: 127–153.

Kogut, B. 1986. On designing contracts to enforce contractibility: Theory and evidence from East-West trade. *Journal of International Business Studies* 17: 47–61.

Marin, D. 2002. Trust versus illusion: What is driving demonetization in Russia? *Economics of Transition* 10: 173–200.

Marin, D., and M. Schnitzer. 1995. Tying trade flows: A theory of countertrade with evidence. *American Economic Review* 85: 1047–1064.

Marin, D., and M. Schnitzer. 2002a. *Contracts in trade and transition: The resurgence of barter*. Cambridge: MIT Press.

Marin, D., and M. Schnitzer. 2002b. The economic institution of international barter. *Economic Journal* 112: 293–316.

Marin, D., and M. Schnitzer. 2005. Disorganization and financial collapse. *European Economic Review* 47: 387–408.

Miras, R., and B. Yeung. 1987. Countertrade and foreign exchange shortages: A preliminary assessment. *Weltwirtschaftliches Archiv* 123: 535–544.

Miras, R., and B. Yeung. 1993. Why countertrade? An economic perspective. *International Trade Journal* 7: 409–433.

C

OECD (Organization for Economic Cooperation and Development). 1981. *East–West trade: Recent developments in countertrade*. Paris: OECD.
OECD. 1985. *Countertrade: Developing countries practices*. Paris: OECD.
Parsons, J.E. 1987. Forms of GDR economic cooperation with the nonsocialist countries. *Comparative Economic Studies* 29: 7–18.

Countervailing Power

Christopher M. Snyder

Keywords
Antitrust; Countervailing power; Mergers; Monopoly; Oligopoly; Pharmaceutical industry; Returns to scale

JEL Classifications
L13

'Countervailing power' is a term coined by J.K. Galbraith (1952) to describe the ability of large buyers in concentrated downstream markets to extract price concessions from suppliers. Galbraith saw countervailing power as an important force offsetting suppliers' increased market power arising from the general trend of increased concentration in US industries. He provided examples such as a nationwide grocery chain extracting wholesale price discounts from food producers, and large auto manufacturers extracting price discounts from steel producers.

The concept of countervailing power was controversial in Galbraith's day (see Stigler's 1954, criticism), and continues to be so today. Formalizing the concept is difficult because it is difficult to model bilateral monopoly or oligopoly, and there exists no single canonical model. Whether and how wholesale discounts to large downstream firms are passed through to final-good consumers is unclear. The concept has the controversial antitrust implication that horizontal mergers between downstream firms may be pro-competitive.

There are a number of theories explaining why large buyers obtain price discounts from sellers. A simple theory is that the cost of serving large buyers is lower per unit than that of serving small buyers. Serving large buyers may involve lower distribution costs. For example, the supplier may be able to ship its product to a large buyer's central warehouse rather than having to ship it to the individual retail outlets owned by small buyers. Serving large buyers may also involve lower production costs. For example, if the supplier's production function exhibits increasing returns to scale and the supplier serves one buyer at a time each production period, per-unit production costs will be lower when serving a large buyer.

Other theories involve more subtle strategic effects. A literature including Horn and Wolinsky (1986), Stole and Zwiebel (1996), Chipty and Snyder (1999), Inderst and Wey (2003) and Raskovich (2003) considers a model in which a monopoly supplier bargains under symmetric information separately and simultaneously with each of a number of buyers. Each buyer regards itself as marginal, conjecturing that all other buyers consummate their negotiations with the supplier efficiently. If aggregate surplus across all negotiations is concave in quantity, the marginal surplus from a transaction involving a large quantity is higher per unit than that from one involving a small quantity. This higher per-unit marginal surplus for large buyers translates into a lower per-unit price. The aggregate surplus function would be concave, for example, if the supplier has increasing marginal production costs. Even if the supplier's cost function were linear, the total surplus function effectively becomes concave if the supplier is assumed to be risk averse, as in Chae and Heidhues (2004) and DeGraba (2005).

Size discounts also emerge if large buyers' outside options are better. In Katz (1987) and Sheffman and Spiller (1992), for example, the larger the buyer, the more credible are its threat of integrating backward and producing the good itself. Size discounts also emerge if the supplier's

outside option is worse when facing a large buyer. In Inderst and Wey (2007), for example, if bargaining with a large buyer breaks down, it is difficult for the supplier to unload this large quantity on the other buyers since this involves marching down these other buyers' declining marginal surplus functions.

Size discounts also emerge if one departs from the bargaining model with a monopoly supplier and instead considers competing suppliers. In Snyder (1998), collusion is difficult to sustain in the presence of a larger buyer because the benefit from undercutting and supplying the buyer is greater. To prevent undercutting in equilibrium, suppliers collude on a lower price for large buyers. In Dana (2004) and Inderst and Shaffer (2007), by pooling their demands and buying as a group from one supplier, buyers can increase the intensity of competition among suppliers of differentiated products.

Several papers have begun to examine the question of whether a downstream firm's countervailing power translates into lower final-good prices, using a model with competing downstream firms (Dobson and Waterson 1997; von Ungern-Sternberg 1996; Chen 2003). This work suggests that an increase in countervailing power can have the opposite effect, raising consumer prices and/or lowering social welfare.

Early empirical studies of countervailing power (see Scherer and Ross 1990, for a survey) took the standard structure–conduct–performance regressions (regressions of supplier profits or markups on supplier concentration using cross-sectional observations at the industry level) and added a buyer-concentration variable, often finding a significantly negative coefficient. Later intra-industry studies found more nuanced circumstances under which buyer-size discounts emerge. Ellison and Snyder (2002) and Sorensen (2003) observed size discounts in pharmaceutical and hospital-services markets only if there were competing, not monopoly, suppliers. In an experimental study, Normann et al. (2007) observed buyer-size discounts only when the total surplus function exhibited a certain curvature, consistent with theory.

See Also

- ▶ Bargaining
- ▶ Galbraith, John Kenneth (1908–2006)
- ▶ Monopsony
- ▶ Price Discrimination (Theory)

C

Bibliography

Chae, S., and P. Heidhues. 2004. Buyers' alliances for bargaining power. *Journal of Economics and Management Strategy* 13: 731–754.

Chen, Z. 2003. Dominant retailers and the countervailing power hypothesis. *RAND Journal of Economics* 34: 612–625.

Chipty, T., and C.M. Snyder. 1999. The role of firm size in bilateral bargaining: A study of the cable television industry. *The Review of Economics and Statistics* 81: 326–340.

Dana, J. 2004. *Buyer groups as strategic commitments*. Mimeo: Northwestern University.

DeGraba, P. 2005. Quantity discounts from risk averse sellers. Working Paper No. 276, U.S. Federal Trade Commission.

Dobson, P.W., and M. Waterson. 1997. Countervailing power and consumer prices. *Economic Journal* 107: 418–430.

Ellison, S.F., and C.M. Snyder. 2002. *Countervailing power in wholesale pharmaceuticals*. MIT: Mimeo.

Galbraith, J.K. 1952. *American capitalism: The concept of countervailing power*. Boston: Houghton Mifflin.

Horn, H., and A. Wolinsky. 1986. Bilateral monopolies and incentive for merger. *RAND Journal of Economics* 19: 408–419.

Inderst, R., and G. Shaffer. 2007. Retail mergers, buyer power, and product variety. *Economic Journal* 117: 45–67.

Inderst, R., and C. Wey. 2003. Bargaining, mergers, and technology choice in bilaterally oligopolistic industries. *RAND Journal of Economics* 34: 1–19.

Inderst, R., and C. Wey. 2007. Buyer power and supplier incentives. *European Economic Review* 51: 647–667.

Katz, M.L. 1987. The welfare effects of third degree price discrimination in intermediate goods markets. *American Economic Review* 77: 154–167.

Normann, H.-T., B.J. Ruffle, and C.M. Snyder. 2007. Do buyer-size discounts depend on the curvature of the surplus function? Experimental tests of bargaining models. *RAND Journal of Economics* 38: 747–767.

Raskovich, A. 2003. Pivotal buyers and bargaining position. *Journal of Industrial Economics* 51: 405–426.

Scherer, F.M., and D. Ross. 1990. *Industrial market structure and economic performance*. Boston: Houghton Mifflin.

Sheffman, D.T., and P.T. Spiller. 1992. Buyers' strategies, entry barriers, and competition. *Economic Inquiry* 30: 418–436.
Snyder, C.M. 1998. Why do large buyers pay lower prices? Intense supplier competition. *Economics Letters* 58: 205–209.
Sorensen, A. 2003. Insurer-hospital bargaining: Negotiated discounts in post-deregulation Connecticut. *Journal of Industrial Economics* 51: 471–492.
Stigler, G.J. 1954. The economist plays with blocs. *American Economic Review* 44: 7–14.
Stole, L.A., and J. Zwiebel. 1996. Organizational design and technology choice under intrafirm bargaining. *American Economic Review* 86: 88–102.
von Ungern-Sternberg, T. 1996. Countervailing power revisited. *International Journal of Industrial Organization* 14: 507–520.

Courcelle-Seneuil, Jean Gustave (1813–1892)

Albert O. Hirschman

Keywords
Advisers; Convertibility; Courcelle-Seneuil, J. G.; Deregulation of banks; Tariffs

JEL Classifications
B31

French economist and economic adviser. Born in the Dordogne, he studied law in Paris, then returned to his native region to manage an industrial firm. At the same time, during the July monarchy, he wrote for Republican newspapers and economic periodicals. After the 1848 revolution, he held briefly a high position in the Ministry of Finance. In the following years he became a frequent contributor to the *Journal des économistes,* and published a successful textbook on banking in 1852. In 1853, the Chilean government contracted him to teach economics at the University of Chile in Santiago, and to be available as official economic adviser; he stayed for ten years, until 1863, when he returned to France. While in Chile he published his most ambitious work in economics, the *Traité théorique et pratique d'économie politique* (1858), which the Chilean government arranged to bring out in a Spanish translation. After his return to France, he resumed his activity as prolific writer of books and articles on economic affairs. He also published several works on political and historical topics and translated into French John Stuart Mill's *Principles of Political Economy,* Summer Maine's *Ancient Law* and William Graham Sumner's *What Social Classes Owe to Each Other.* He was appointed councillor of state in 1879, and three years later was elected member of the Académie des Sciences Morales et Politiques.

Throughout his life, Courcelle-Seneuil was a stalwart defender of free trade and laissez-faire. Charles Gide, the co-author (with Charles Rist) of a well-known history of economic doctrines, wrote about him in rather sarcastic terms:

> He was virtually the *pontifex maximus* of the classical school; the holy doctrines were entrusted to him and it was his vocation to denounce and exterminate the heretics. During many years he fulfilled this mission through book reviews in the *Journal des économistes* with priestly dignity. Argus-eyed, he knew how to detect the slightest deviations from the liberal school. (Gide 1895, p. 710)

Courcelle-Seneuil's special interest, starting with the publication of a small book on bank reform in 1840, was the introduction of more freedom into banking or, to use a modern term, the 'deregulation' of this industry. Above all, he advocated the abolition of the Bank of France's exclusive right of issue. According to Gide, Courcelle-Seneuil was more esteemed in England and the United States than in France. In any event, adoption of his monetary and banking proposals was never seriously considered in his own country.

Once in Chile, Courceile-Seneuil became a powerful policymaker and influential teacher. He arrived at a time when the international prestige of the laissez-faire doctrine was at its height and when gold booms and subsequent busts in California and Australia caused considerable fluctuations in Chile's agricultural exports to these areas, creating a need for flexible short- and long-term

credit facilities. This combination of events, joined with the prestige emanating from the foreign savant, permitted him to obtain in Chile what he had failed to achieve in his own country: under his guidance, the administration of Manuel Montt (1851–1861) promulgated a banking law that established total freedom for any solvent person to found a bank and permitted all banks to issue currency subject only to one limitation: the banknotes in circulation were not to exceed 150 per cent of the issuing bank's capital.

Courcelle-Seneuil's advice was also sought in connection with a new customs tariff and here again he achieved substantial change: the level of protection was severely cut back, although some tariffs were retained for revenue purposes.

But the principal influence exercised by Courcelle-Seneuil resided in his forceful teaching: as the University of Chile's first professor of economics, he was apparently successful in instilling doctrinaire zeal in his students, some of whom later became influential policymakers. Thus, Chilean historians have not only traced the abandonment of convertibility in 1878 to the permissiveness of the 1860 Banking Law and the lack of industrial development to the 1864 tariff; they also see Courcelle-Seneuil's indirect influence in the acquisition of the nitrate mines of Tarapacá by private foreign interests after Chile's victory over Peru in the War of the Pacific (1882) had given it title to the mines. Alienation of the mines was indeed recommended by a government committee dominated by Courcelle-Seneuil's disciples, who felt, like their teacher, that state ownership and management of business enterprises was to be strictly shunned. Secular inflation, industrial backwardness, domination of the country's principal natural resources by foreigners – all of these protracted ills of the Chilean economy have been attributed to the French expert.

Since the economically advanced countries were also those where economic science first flourished, they soon produced a peculiar export product: the foreign economic expert or adviser. Courcelle-Seneuil is probably the earliest prototype of the genre and his ironic career in Chile exhibits characteristics that were to remain typical of numerous later representatives. First, the adviser is deeply convinced that, thanks to the advances of economic science, he knows the correct solutions to economic problems no matter they may arise. Secondly, the country which invites the expert looks forward to his advice as to some magic medicine which will work even when (perhaps especially when) it hurts. Some countries seem particularly prone to this attitude. In Chile foreign or foreign-trained experts have played key roles at crisis junctures, from Courcelle-Seneuil in the mid-19th century to Edwin Kemmerer in the 1920s, the Klein-Saks Mission in the 1950s, and finally to the 'Chicago boys' in the 1970s. Thirdly, the influence of the adviser derives not only from the intrinsic value and persuasiveness of his message, but from the fact that he usually has good connections in his home country and can therefore facilitate access to its capital market. Courcelle-Seneuil, for example, suspended his university courses in 1858–1859 to accompany a Chilean financial mission that travelled to France in search of a railroad construction loan. Fourthly, the foreign adviser is often criticized for wishing to transplant the institutions of his own country to the country he advises, but his real ambition is more extravagant: it is to endow the country with those ideal institutions which exist in his mind only, for he has been unable to persuade his own countrymen to adopt them. Fifthly, history in general, and nationalist historiography in particular, is likely to be unkind to the foreign adviser. In retrospect he can easily become a universal scapegoat: whatever went wrong is attributed to his nefarious influence. This demonization is more damaging than the adviser himself could possible have been: it forestalls authentic learning from past experience.

Selected Works

1840. Le crédit et la Banque. *Paris.*
1858. Traité théorique et pratique d'économie politique, *2 vols. Paris: Amyot.*
1867. *La Banque libre.* Paris: Guillaumin.

Bibliography

Encina, F. 1951. *Historia de Chile, vol. 18, ch. 58*. Santiago: Nascimiento.
Fuentealba, H.L. 1946. *Courcelle-Seneuil en Chile: Errores del liberalismo económico*. Santiago: Prensas de la Universidad de Chile.
Gide, C. 1895. Die neuere volkswirtschaftliche Litteratur Frankreichs. *Schmollers Jahrbuch*.
Hirschman, A.O. 1963. *Journeys toward progress*, 163–168. New York: Twentieth Century Fund.
Journal des économistes. 1892. Obituary [of M.J.G. Courcelle-Seneuil]. July.
Juglar, C. 1895. Notice sur la vie et les travaux de M.J.G. Courcelle-Seneuil. Académie des Sciences Morales et Politiques, *Compte Rendu*, 850–82.
Pinto, S.C. 1959. *Chile, un caso de desarrollo frustrado*. Santiago: Edit. Universitaria.
Will, R.M. 1964. The introduction of classical economics into Chile. *Hispanic- American Historical Review* 44 (1): 1–21.

Cournot Competition

Andrew F. Daughety

Abstract

Cournot's 1838 model of strategic interaction between competing firms has become the primary workhorse for the analysis of imperfect competition, and shows up in a variety of fields, notably industrial organization and international trade. This article begins with a tour of the basic Cournot model and its properties, touching on existence, uniqueness, stability, and efficiency; this discussion especially emphasizes considerations involved in using the Cournot model in multi-stage applications. A discussion of recent applications is provided as well as a reference to an extended bibliography of approximately 125 selected publications from 2001 through 2005.

Keywords

Auctions with competition; Bertrand competition; Best-reply dynamics; Best-response functions; Complementarities; Conjectural variations; Cournot competition; Cournot equilibrium; Cournot models; Cournot, A. A.; Differentiated products; Dynamic stability; Efficiency; Existence theorems; Imperfect competition; Information sharing among firms; Licences; Market power; Mergers; Multiple equilibria; Multi-stage games; Multi-stage models of competition; Networks; Oligopoly; Patents; Product differentiation; Repeated games; Research and development; Signalling; Strategic substitutes; Subgame perfection; Supermodular games; Uniqueness; Welfare

JEL Classifications

D4

The classic Cournot model is static in nature, with each (single-product) firm's strategy being the quantity of output it will produce in the market for a specific homogeneous good; as Kreps (1987) observed, Cournot's model was an early progenitor of Nash's famous paper. Many recent applications have involved multi-stage games; for example, each of n firms might first simultaneously choose investment levels (say, in cost-reducing R&D) and then simultaneously choose output levels in the second stage. Often now used in such a manner, we will see that the Cournot model is doing well, contributing to a range of new research, as it moves towards the two-century mark.

The Basic One-Stage Model and Associated Concepts

Consider an industry comprised of n firms, each firm choosing an amount of output to produce. Firm i's *output level* is denoted as q_i, $i = 1, \ldots, n$; let the vector of firm outputs be denoted $\boldsymbol{q} \equiv (q_1, q_2, \ldots, q_n)$. The firms' products are assumed to be perfect substitutes (the *homogeneous-goods* case); let Q denote the aggregate industry output level (that is, $Q \equiv \sum_{i=1}^{n} q_i$). We will refer to the $(n-1)$ vector of output levels chosen by firm i's

rivals as $\boldsymbol{q}_{-i}$; so, let ($\boldsymbol{q}_{-i}$, q_i) also be the n-vector $\boldsymbol{q}$. Market demand for the perfect-substitutes case is a function of aggregate output and its inverse is denoted as $p(Q)$; furthermore, let firm i's cost of producing q_i be denoted as $c_i(q_i)$. Thus, firm i's *profit* function is written as $\pi^i(\boldsymbol{q}) \equiv p(Q)q_i - c_i(q_i)$. All elements of the model are assumed to be commonly known by the firms, though extensions allowing incomplete information are not uncommon.

A *Cournot equilibrium* consists of a vector of output levels, $\boldsymbol{q}^{CE}$, such that no firm wishes to unilaterally change its output level when the other firms produce the output levels assigned to them in the (purported) equilibrium. Alternatively put (and reversing history), it is a Nash equilibrium of the normal-form game with quantities as strategies chosen from a compact space (for example, q_i in [0, Q^*], for some appropriate Q^*, such as $p(Q^*)= 0$) and with the $\pi^i(q)$ as the payoff functions. Thus, q^{CE} is a Cournot equilibrium if the following n equations are satisfied:

$$\pi^i\left(\boldsymbol{q}^{CE}\right) \geq \pi^i\left(\boldsymbol{q}^{CE}_{-i}, q_i\right) \text{ for all values of } q_i,$$
$$\text{for i} = 1, \ldots, n.$$

In analysing his model applied to a duopoly (he also considered the n-firm version), Cournot provided the notion of *best-response functions*. In the duopoly case, this is a pair of functions, $\psi^1(q_2)$ and $\psi^2(q_1)$, which provide the profit-maximizing choice of output for firm 1 and 2 (respectively), given conjectures about the output level chosen by the rival firm (that is, each firm's choice of its output level reflects a *best-response property*). Hence, $\psi^i(q_j)$= arg $\max_q \pi^i(q, q_j)$; $i, j = 1, 2$; $i \neq j$. That is, we want $\psi^i(q_j)$ to be the solution to firm i's first-order condition: $p(\psi^i(q_j) + q_j) + p'(\psi^i(q_j) + q_j)\psi^i(q_j) - c'(\psi^i(q_j)) = 0$, $i, j = 1, 2$, $i \neq j$. We ll assume for now that the problem has a nice solution and that some sort of sufficiency condition holds (for example, strict quasi-concavity of profits), but the discussion below on existence and uniqueness of equilibrium shows that such classical assumptions are overly strong and are overly restrictive for some modern applications, such as those involving multi-stage games or discontinuous cost functions. More generally, $\psi^i(\mathrm{q_j})$ could be a correspondence (a point-to-set map); we generally restrict the discussion below to functions, and assume as much differentiability as needed.

If output-level choices are best responses to conjectures about each firm's rival's choice of output, *and* if these conjectures are correct in equilibrium, then the resulting vector of output levels provides a Cournot equilibrium: $q_i^{CE} = \psi^i\left(q_j^{CE}\right)$ for $i,j = 1, 2$, and $j \neq i$. In other words, the equilibrium occurs where the best-response functions cross when graphed in the space of output levels. Generalizing to n firms, this condition can be written as $q_i^{CE} = \psi^i\left(q_j^{CE}\right)$ for $i = 1, \ldots, n$: $\boldsymbol{q}^{CE}$ is a Cournot equilibrium if it consists of mutual best-responses for all the firms.

Some variations on the basic model are worth mentioning. If the cost function for a firm has both fixed and variable components, and if the fixed component is avoidable (that is, is zero at zero output), then the best-response function for the firm will be discontinuous at the positive output level where variable profits just cover the avoidable cost. This is important for two reasons. First, avoidable fixed costs are not unusual in many entry scenarios: think of an airline entering a market where there are already some competitors, with the avoidable cost being advertising. Second, this discontinuity could mean that the only equilibrium might involve some or all firms choosing to not enter (or to exit) the market, even if absent these avoidable costs $\boldsymbol{q}^{CE}$ would be strictly positive.

Another avenue for interaction would consider imperfect factor markets, so that instead of $c_i(q_i)$ the cost function for firm i would be written as $c_i(\boldsymbol{q}_{-i}, q_i)$; then strategic interaction occurs not only through revenue but also via factor markets. Finally, if the model is one of short-run competition, then the output level of the firm may be restricted to be less than some predetermined capacity level; a simple version is that there are parameters k_i, $i = 1, \ldots, n$, such that a constraint on firm i's quantity choice is $q_i \leq k_i$; $i = 1, \ldots, n$; this induces a vertical segment (at the capacity level) in a firm's best-response function. Such capacity levels might be choices made in an earlier stage.

Finally, a number of papers develop 'non-Cournot' models which generate Cournot-model results. Kreps and Scheinkman (1983) provide a two-stage model of capacity choice followed by price setting in a homogeneous-goods duopoly; the result is a unique subgame-perfect equilibrium with Cournot capacities and a market-clearing price consistent with the standard Cournot model (however, Davidson and Deneckere 1986, show that this result is especially sensitive to the basis for rationing consumers over firms when out-of-equilibrium firm-level demand exceeds capacity). Klemperer and Meyer (1986) analyse a one-stage game wherein duopolists producing heterogeneous goods non-cooperatively choose either a price or a quantity as the firm's strategy; under either multiplicative or additive error in the demand function, if marginal costs are upward sloping, the outcome is that predicted by the Cournot model (applied to the heterogenous-goods case; see the discussion of this case in Section "Properties of the Cournot equilibrium" below). The classic embedding of the Cournot model is that of Bowley (1924), the best-known developer of models with 'conjectural variations' (CV). This is a static story wherein the first-order conditions in the analysis include firm i's conjecture of each rival's reaction to a small change in firm i's quantity (for example, $\partial_{qj}/\partial_{qj}$ need not be zero for each $j \neq i$); different values of the CV generate competitive, collusive, or Cournot outcomes (among others). Such a handy static embedding of alternative degrees of competition has been employed in a number of theoretical applications, and in a variety of empirical analyses trying to estimate market power. However, Daughety (1985) shows that a basic rationality requirement (that each firm's CV be the same as the actual slope of the best-response function) leads to the Cournot outcome, so that alternative CV values violate this form of rational expectations. Furthermore, Korts (1999) shows that empirical analyses using the CV approach to assess market power will generally mis-measure the degree of competitiveness of the industry.

Properties of the Cournot Equilibrium

For most of this section we emphasize results for an n-firm, homogeneous-goods, complete-information model, where a firm's cost function depends only on that firm's output level. As suggested earlier, possibly one of the most important reasons for the continuing interest in the properties of the Cournot equilibrium is that Cournot competition is frequently used as a final stage in a variety of models; analysis employing such refinements as subgame perfection rely on a well-behaved subgame.

Existence, Uniqueness and Stability

Novshek (1985) provides an existence theorem that has quite practical uses (for expository purposes we consider a slightly less general version). Besides continuity and twice differentiability of the inverse demand function, $p(Q)$, Novshek's existence theorem requires that: (1) $p(Q)$ crosses the quantity axis at a finite value and is strictly decreasing for quantities below that cut point; (2) the marginal revenue for each firm is decreasing in the aggregate output of its rivals; and (3) each firm's cost function is non-decreasing and lower semi-continuous. Requirement (2) is written formally as $p'(Q_{-i} + q_i)\ p''(Q_{-i} + q_i)\ q_i < 0$, where $Q_{-i} \equiv Q - q_i$, for all i. This is equivalent to the assumption that $\partial^2\pi^i(\boldsymbol{q})/\partial Q_{-i}\partial_{qi} < 0$ for all i, that is, that Q_{-i} and q_i are *strategic substitutes*, which means that an expansion in Q_{-i} implies that the optimal q_i falls. The third requirement means that costs cannot fall as the output level is increased and that cost functions can have jumps (discontinuities), as long as the functions are continuous from the left. This was a substantial improvement over previous existence theorems and it allows for an important case: avoidable fixed costs, such as those in the airline-entry example mentioned earlier. Amir (1996) applies an ordinal version of the theory of *supermodular games* to the existence issue (see Vives 2005, for a recent survey of supermodular games; see also Amir 2005, for a comparison of *ordinal* and *cardinal complementarity* in this

context); this change of techniques allows for weaker demand conditions (primarily that log $p(Q)$ is concave) but requires a slightly stronger condition on each firm's cost function (marginal costs are positive, so models wherein marginal costs might be zero – as might occur with capacity competition – are left out) in order to guarantee that a Cournot equilibrium exists. As an example of the advantages concerning demand analysis, let $p(Q) = \left(Q - \overline{Q}\right)^2$ for $Q \leq \overline{Q}$, and zero otherwise. Such a function satisfies (1) above, is log-concave (actually, convex), but is excluded from consideration by Novshek's second condition.

Gaudet and Salant (1991) provide conditions for a Cournot equilibrium to be *unique* which address an important consideration when Cournot models are used in a subgame of a larger game: their theorem allows for degeneracy (one or more firms produce zero output but have marginal cost equal to the equilibrium price); thus, such firms are just at the shutdown point in the equilibrium. In a one-stage application this could be eliminated via a small perturbation in the parameters, but in a multi-stage application such an outcome need not be pathological, as some of the second-stage 'parameters' are strategic variables in the first-stage model (the authors provide a simple, full-information entry game to illustrate this). The sufficient conditions for uniqueness are (not surprisingly) more restrictive than those for existence (on the assumption that Novshek's conditions hold as well): (1) each firm's cost function must be twice continuously differentiable and strictly increasing; and (2) the slope of the marginal cost function is strictly bounded above the slope of the demand function. Thus, concave costs are allowed, to some degree, but the cost function cannot be 'too concave', even on subsets of its domain.

Cournot provided an explicit dynamic stability argument for his model by imagining sequential play by each agent (myopically best-responding in the current period to the existing output levels of all rivals); this is referred to as *best-reply dynamics* and when this process converges the solution is termed *stable*. Using best-reply dynamics to rationalize a static solution has, historically, been a source of substantial criticism, but nonetheless some papers use the requirement of Cournot stability to select an equilibrium when there are multiple equilibria (dynamic stability should not be confused with equilibrium refinement criteria in game theory such as strategic stability). A sufficient condition in the duopoly case is that $|\partial\psi^1(q_2)/\partial q_2||\partial\psi^2(q_1)/\partial q_1| < 1$ (see Fudenberg and Tirole 1991); see Seade (1980) for more general conditions (and problems) for best-reply dynamics in the n-firm case. For an approach employing an explicit evolutionary process via replicator dynamics with noise, with firms able to choose 'behavioural' strategies (including, but not limited to, best-reply), see Droste et al. (2002).

Welfare

Two types of inefficiency can occur in a Cournot equilibrium: the equilibrium price exceeds the marginal cost of production, and aggregate output is inefficiently distributed over the firms. Compare the first-order conditions for firms in a duopoly, each producing under conditions of non-decreasing marginal costs (that is, $p(Q) + p'(Q)q_i = c_i'(q_i), \quad i = 1, 2$) with those for a central planner choosing q_1 and q_2 so as to maximize total surplus: $p(Q) = c_i'(q_i), \quad i = 1, 2$. Clearly, if demand is downward-sloping at the equilibrium, aggregate output in the Cournot equilibrium will be less than what the social planner would choose. However, a second distortion can be seen in this comparison: under the social planner, each firm's marginal costs are equalized with the others'. This will hold only in a symmetric Cournot equilibrium (where $q_1 = q_2$): production is, in general, inefficiently allocated across the firms.

The maldistribution of production implies that strategic interaction readily may yield counter-intuitive welfare results. As a simple example, consider a duopoly wherein (inverse) industry demand is $p = a - Q$ and firm i's cost function is $c_i(q) = C_i q$, $i = 1, 2$, with $a > C_1 > C_2 > 0$; that is, the linear demand, constant- but-unequal-marginal-cost case. It is straightforward to find the equilibrium and show that it is interior and

unique. Let W be the sum of producers' and consumers' surplus. Then a little work shows that $dW/dC_1 > 0$ if $11C_1 - 7C_2 - 4a > 0$; to see that these conditions are non-empty, consider the parameter specification ($a = 20$, $C_1 = 13$, $C_2 = 8$), which satisfies all the foregoing requirements. The point of the example is that a *reduction* of firm 1's marginal cost leads to a *decrease* in equilibrium welfare. Thus, strategic interaction by the firms in the marketplace can lead to reversals of the usual welfare intuition that cost-improving technological change is beneficial. The reason this occurs is that the cost reduction results in an increase in the high-cost firm's equilibrium output level and a (smaller) decrease in the low-cost firm's output level; this increased inefficiency in aggregate production can be sufficient to overwhelm other efficiency improvements (such as the increase in industry output). This is similarly true if in the above model firm 2 is an incumbent monopolist (using simple monopoly pricing) and firm 1 an entrant: welfare will fall due to entry.

In the n-firm version of the constant-marginal-cost model, changes in the distribution of production costs (holding the mean fixed) do not affect industry output; this is seen by summing over the first-order conditions, whence $np(Q) + p'(Q)Q = \sum_{i=1}^{n} C_i$. Bergstrom and Varian (1985) showed that (on the assumption that the pre-and post-change equilibria are interior) such mean-preserving changes in the marginal costs strictly improve welfare if and only if the variance of the marginal costs strictly increases; the reason is that the aggregate cost of production has decreased if the variance increases. Salant and Shaffer (1999) extend this idea to consider the effects of changes in first-stage parameters (for example, cost-reducing R&D investments) on second-stage costs in models wherein Cournot competition is employed in the second stage. They argue that, since aggregate production costs are *maximized* when all firms have the same costs, it is the asymmetric equilibria in such games (which are often assumed away) which may yield the most important outcomes to examine, from both a social and a private perspective.

Does entry necessarily reduce the equilibrium price? A recent contribution provides a clean result if we restrict attention to the symmetric case wherein all firms have the same twice continuously differentiable and non-decreasing cost function, and demand is continuously differentiable and downward-sloping. Amir and Lambson (2000) show that the equilibrium price falls with an increase in the number of competitors if, for all Q, $p'(Q) < c''(q)$ for all q in [0, Q]. Thus, even with some degree of returns to scale (for example, as might occur with U-shaped average costs), entry will reduce price, at least with identical firms. However, Hoernig (2003) shows that, even if the equilibria are stable and there are no returns to scale, price can rise with entry if products are differentiated.

If the products of the firms are imperfect substitutes (that is, products are differentiated), then (in general) there is no aggregate demand function $p(Q)$; rather firm i's inverse demand function would be written as $p_i(\boldsymbol{q})$ and profits would be written as $\pi^i(\boldsymbol{q}) = p_i(\boldsymbol{q})q_i - c_i(q_i)$, $i = 1, \ldots, n$. Welfare in this model can be contrasted with a reformulation of the model so that each firm chooses a price for its product; standard parlance is to call the price-strategy model the *(differentiated products) Bertrand model* (even though Bertrand's famous review of Cournot did not envision heterogeneity in products; see Friedman's 1988 translation of Bertrand's review). Without going into detail on the (differentiated products) Bertrand model, Singh and Vives (1984) have shown (for linear, symmetric demand and constant marginal costs in a duopoly setting) that, while profits under Cournot competition exceed those under Bertrand competition, total surplus is higher under Bertrand competition than under Cournot competition. Note that this result holds in the one-stage game. However, these results may be reversed in a two-stage application. For example, Symeonidis (2003) considers R&D investment with spillovers in a two-stage game, and shows that (at least for a portion of the parameter space) Cournot competition leads to higher welfare than Bertrand competition. The basic intuition is that, if profits are higher for second-stage Cournot competition

than for second-stage Bertrand competition, and first-stage investment is inefficiently low in either case, then the increased second-stage profits may partly correct the inefficiently low first-stage investment, leading to an overall welfare gain for competition in quantities rather than prices.

Finally, convergence of a Cournot equilibrium to a competitive equilibrium, as the number of firms grows, was considered by Cournot in Chapter ▶ "Hearn, William Edward (1826–1888)" of his book, and has been the subject of a number of papers; see Novshek and Sonnenschein (1978, 1987) for a general equilibrium treatment where appropriate replication of Cournot economies yields equilibria arbitrarily close to the Walrasian equilibrium; see Alos-Ferrer (2004) for an evolutionary model (which allows for memory) at the level of an industry.

Applications

The literature exploring and applying the Cournot model is vast; an earlier extended bibliography can be found in Daughety (1988/2005). The more recent literature employing the Cournot model is already becoming significant in size: a survey of articles in 16 top mainline and field journals, for the period 2001–5, netted approximately 125 articles exploring or applying the Cournot model in one of its various common forms. An online Excel file of (abbreviated) citations and some characteristics of each article (number of firms, number of stages, welfare considerations, informational regime, and topic classification), as accessed on 21 November 2006, is available at http://www.vanderbilt.edu/Econ/faculty/Daughety/ExtendedCournotBib2001-2005.xls

However, some excellent papers have undoubtedly been missed (not to mention papers from the 1990s), and space limitations preclude anything beyond the briefest of tours and just a taste of the literature, so only a very few can be discussed below. This section addresses five topics which account for a significant portion of the literature, three areas that overlap other fields, and two (comparatively) new areas of research.

Delegation

Vickers (1985) uses an *n*-firm, two-stage model to examine performance measures for managers. Restricting the manager's performance measure to be a weighted average of profits and output, with the weights determined by the owner of each firm in the first stage, he shows that the weight on output is non-zero. This makes each manager more aggressive (each chooses to produce a higher output level), thereby leading to lower profits per firm. Sklivas (1987) considers the differentiated-products Bertrand version and shows that owners choose weights on revenue and profits so as to make managers more passive (they post higher prices), leading to increased profits. Miller and Pazgal (2001) have unified this literature, showing that incentive schemes based on own and rival's profits result in an equilibrium which is insensitive to whether the firm chooses price or quantity as its strategic variable.

Information Transfer

Vives (1984), Gal-Or (1985), and Li (1985) all consider variants of 'information transfer' models to examine the possibility of information sharing, whereby firms may choose to pool information on either demand or cost parameters. These models are analysed as *Bayesian–Nash games*, so that, before seeing a private signal about the parameter of interest (for example, the demand intercept), each firm chooses whether or not to share the information with the other firms; then information is received and production (or pricing) occurs in the second stage. The nature of the good (substitutes or complements), the type of information (common or individual), and the strategy space (quantities or prices) all affect whether firms will share information. Ziv (1993) relaxes the verifiability of information and finds that firms will send misleading information if they can; he then considers mechanisms for eliciting truthful messages.

Intellectual Property

Katz and Shapiro (1985) and Kamien and Tauman (1986) consider the licensing of innovations in an oligopoly. Katz and Shapiro employ a three-stage duopoly game in which the innovation is

developed, then a single license is auctioned, and then the firms compete. Kamien and Tauman use a two-stage, *n*-firm game with a posted price for the innovation (a fee or a royalty), followed by competition. More recently, Fauli-Oller and Sandonis (2003) consider optimal competition policy when considering licences as an alternative to merger. Anton and Yao (2004) allow for weak patent protection and consider how disclosure of information about an innovation (for example, through the patent application) can be a signalling device to influence competitors, but those same competitors may be able to employ the information to successfully use (infringe on) the patent; here small innovations are patented and substantial innovations are protected through secrecy.

Mergers

Salant et al. (1983) show that exogenously determined mergers of a subset of firms in the constant-marginal-cost set-up yields a problematical result: a sufficient condition for a merger to be unprofitable is that it involve less than 80 per cent of the industry, hardly a resounding endorsement of using such a model to analyse mergers. This result, however, is partly driven by the assumptions of homogeneous products, constant unit costs, and industry structure. Perry and Porter (1985) show that various mergers can be profitable if firms have sufficiently increasing marginal costs. Daughety (1990), using a two-tiered-industry, *n*-firm model, with m firms choosing output in the first stage (tier) and $n - m$ firms choosing output in the second stage, shows that if $1 < m < n$, then, when m is comparatively small ($m < n/3$), mergers of two second-tier firms to make a first-tier firm can be both profitable and social-welfare-enhancing, even though such mergers increase concentration and have no cost synergies (all firms have identical constant unit costs). Recently, Pesendorfer (2005), using a repeated game model with entry, has found that merger to monopoly may not be profitable, but merger in a non-concentrated industry can be; these differences from the previous literature partly reflect long-run versus short-run profitability computations.

R&D

D'Aspremont and Jacquemin (1988) considered cost-reducing R&D in the presence of spillovers, and considered both non-cooperative and cooperative R&D decision-making; there have been a number of recent papers on cost-reducing spillovers (see, for example, Zhao 2001, for more on the negative welfare effects of cost-reducing innovation, and Symeonidis 2003, cited in Section "Properties of the Cournot equilibrium" above, as well as the work discussed below under the subject of auctions with competition). Toshimitsu (2003) considers the incentive and welfare properties of quality-based R&D subsidies for firms in a model of endogenously determined product quality (and thus product differentiation); subsidizing high quality is welfare-enhancing (independent of whether the Cournot or Bertrand model is employed).

Other Fields

Areas of ongoing effort which extend into other fields include *experimental economics*, *the financial structure of the firm* (see, for example, Brander and Lewis 1986, on determinate debt-equity due to imperfect competition, and see Povel and Raith 2004, extending Brander and Lewis via endogenously determined debt contracts); and *international trade* (see, for example, Brander and Spencer 1985, analysing the strategic use of subsidies in international competition; Mezzetti and Dinopoulos 1991, discussing domestic firm–union bargaining and import competition; and Spencer and Qiu 2001, concerning relationship-specific investments and trade).

New Topics

Finally, a few examples of comparatively new topics. While auctions with private information has long been an area of interest, the developing literature on *auctions with competition* has started to take seriously the combination of incomplete information and post-auction competition. For example, see Das Varma (2003) or Goeree (2003), who find that signalling by winners of an auction causes bids to be biased when post-auction interaction between the auction's winner and losers can be influenced by the size of the bid.

A nice example is when firms have private information about how acquiring a cost-reducing innovation might affect the firm's production costs, and bidding for a licence for the innovation precedes Cournot oligopoly interaction; here signalling with a high bid suggests that the winner will have low costs and will produce a high level of output.

A second new area is *networks*; one recent example is Goyal and Moraga-Gonzalez (2001), who model bilateral agreements to share knowledge, and allow for the possibility of partial collaboration, via considering possible networks of relationships. They examine how the nature of the firms' interaction in markets can contribute to the instability of certain types of strategic alliances and the stability of other ones.

A Broader Perspective on Cournot Competition

If alive to critique this essay, Cournot might view the interpretation of the term 'Cournot competition' being limited merely to the legacy of his oligopoly analysis to be an overly restrictive interpretation of the assignment. And well he should. Hicks (1935, 1939) argues that Cournot was the first to present a modern model of monopoly as well as the precise conditions for perfect competition; furthermore, as noted earlier, Cournot's eighth chapter concerned 'unlimited competition'. In the 1937 Cournot Memorial session of the Econometric Society, A. J. Nichol (1938) observed that, if ever there was an apt illustration of Carnegie's dictum that 'It does not pay to pioneer', then Cournot's life and work would be it. Cournot's oligopoly model was essentially ignored for many years, or was relegated to dusty corners of microeconomics texts, but over recent decades it has come to be an essential tool in many an economist's toolbox, and is likely to continue as such.

See Also

- ▶ Bertrand Competition
- ▶ Experimental Economics

Bibliography

Alos-Ferrer, C. 2004. Cournot versus Walras in dynamic oligopolies with memory. *International Journal of Industrial Organization* 22: 193–217.

Amir, R. 1996. Cournot oligopoly and the theory of supermodular games. *Games and Economic Behavior* 15: 132–148.

Amir, R. 2005. Ordinal versus cardinal complementarity: The case of Cournot oligopoly. *Games and Economic Behavior* 53: 1–14.

Amir, R., and V.E. Lambson. 2000. On the effects of entry in Cournot markets. *Review of Economic Studies* 67: 235–254.

Anton, J.A., and D.A. Yao. 2004. Little patents and big secrets: Managing intellectual property. *RAND Journal of Economics* 35: 1–22.

Bergstrom, T.C., and H.R. Varian. 1985. Two remarks on Cournot equilibria. *Economics Letters* 19: 5–8.

Bertrand J. 1883. Review of Walras's *Théorie mathématique de la richesse social* and Cournot's *Recherches sur les principes mathématiques de la théorie des richesses*. Trans. J.W. Friedman, in A.F. Daughety (1988).

Bowley, A.L. 1924. *The mathematical groundwork of economics*. Oxford: Oxford University Press.

Brander, J.A., and T.R. Lewis. 1986. Oligopoly and financial structure: The limited liability effect. *American Economic Review* 76: 956–970.

Brander, J.A., and B. Spencer. 1985. Export subsidies and international market share rivalry. *Journal of International Economics* 18: 83–100.

Cournot, A. 1929. *1838. Researches into the mathematical principles of the theory of wealth. Trans. N.T Bacon*. New York: Macmillan.

d'Aspremont, C., and A. Jacquemin. 1988. Cooperative and noncooperative R&D in duopoly with spillovers. *American Economic Review* 78: 1133–1137.

Das Varma, G. 2003. Bidding for a process innovation under alternative modes of competition. *International Journal of Industrial Organization* 21: 15–37.

Daughety, A.F. 1985. Reconsidering Cournot: The Cournot equilibrium is consistent. *RAND Journal of Economics* 16: 368–379.

Daughety, A.F. 1988. *Cournot oligopoly – characterization and applications*. New York: Cambridge University Press (reprinted 2005).

Daughety, A.F. 1990. Beneficial concentration. *American Economic Review* 80: 1231–1237.

Davidson, C., and R. Deneckere. 1986. Long-run competition in capacity, short-run competition in price, and the Cournot model. *RAND Journal of Economics* 17: 404–415.

Droste, E., C. Hommes, and J. Tunistra. 2002. Endogenous fluctuations under evolutionary pressure in Cournot competition. *Games and Economic Behavior* 40: 232–269.

Fauli-Oller, R., and J. Sandonis. 2003. To merge or to license: Implications for competition policy. *International Journal of Industrial Organization* 21: 655–672.

Fudenberg, D., and J. Tirole. 1991. *Game theory*. Cambridge, MA: MIT Press.
Gal-Or, E. 1985. Information transmission – Cournot and Bertrand. *Review of Economic Studies* 53: 85–92.
Gaudet, G., and S. Salant. 1991. Uniqueness of Cournot equilibrium: New results from old methods. *Review of Economic Studies* 58: 399–404.
Goeree, J.K. 2003. Bidding for the future: Signaling in auctions with an aftermarket. *Journal of Economic Theory* 108: 345–364.
Goyal, S., and J.L. Moraga-Gonzalez. 2001. R&D networks. *RAND Journal of Economics* 32: 686–707.
Hicks, J.R. 1935. Annual survey of economic theory: The theory of monopoly. *Econometrica* 3: 1–12.
Hicks, J.R. 1939. *Value and Capital*. 2nd ed. London: Oxford University Press.
Hoernig, S.H. 2003. Existence of equilibrium and comparative statics in differentiated goods Cournot oligopolies. *International Journal of Industrial Organization* 21: 989–1019.
Kamien, M.I., and Y. Tauman. 1986. Fees versus royalties and the private value of a patent. *Quarterly Journal of Economics* 101: 471–492.
Katz, M.L., and C. Shapiro. 1985. On the licensing of innovations. *RAND Journal of Economics* 16: 504–520.
Klemperer, P., and M. Meyer. 1986. Price competition vs. quantity competition: The role of uncertainty. *RAND Journal of Economics* 17: 618–638.
Korts, K.S. 1999. Conduct parameters and the measurement of market power. *Journal of Econometrics* 88: 227–250.
Kreps, D.M. 1987. Nash equilibrium. In *The new palgrave: A dictionary of economics*, ed. J. Eatwell, M. Milgate, and P. Newman, Vol. 3. London: Macmillan.
Kreps, D.M., and J.A. Scheinkman. 1983. Quantity precommitment and Bertrand competition yield Cournot outcomes. *Bell Journal of Economics* 14: 326–337.
Li, L. 1985. Cournot oligopoly with information sharing. *RAND Journal of Economics* 16: 521–536.
Mezzetti, C., and D. Dinopoulos. 1991. Domestic unionization and import competition. *Journal of International Economics* 31: 79–100.
Miller, N.H., and A.I. Pazgal. 2001. The equivalence of price and quantity competition with delegation. *RAND Journal of Economics* 32: 284–301.
Nichol, A.J. 1938. Tragedies in the life of Cournot. *Econometrica* 3: 193–197.
Novshek, W. 1985. On the existence of Cournot equilibrium. *Review of Economic Studies* 52: 85–98.
Novshek, W., and H. Sonnenschein. 1978. Cournot and Walras equilibrium. *Journal of Economic Theory* 19: 223–266.
Novshek, W., and H. Sonnenschein. 1987. General equilibrium with free entry. *Journal of Economic Literature* 25: 1281–1306.
Perry, M.K., and R.H. Porter. 1985. Oligopoly and the incentive for horizontal merger. *American Economic Review* 75: 219–227.
Pesendorfer, M. 2005. Mergers under entry. *RAND Journal of Economics* 36: 661–679.
Povel, P., and M. Raith. 2004. Financial constraints and product market competition: Ex ante vs. Ex post incentives. *International Journal of Industrial Organization* 22: 917–949.
Salant, S.W., and G. Shaffer. 1999. Unequal treatment of identical agents in Cournot equilibrium. *American Economic Review* 89: 585–604.
Salant, S.W., S. Switzer, and R.J. Reynolds. 1983. Losses from horizontal merger: The effects of an exogenous change in industry structure on Cournot–Nash equilibrium. *Quarterly Journal of Economics* 98: 185–199.
Seade, J. 1980. The stability of Cournot revisited. *Journal of Economic Theory* 23: 15–27.
Singh, N., and X. Vives. 1984. Price and quantity competition in a differentiated duopoly. *RAND Journal of Economics* 15: 546–554.
Sklivas, S.D. 1987. The strategic choice of managerial incentives. *RAND Journal of Economics* 18: 452–458.
Spencer, B.J., and L.D. Qiu. 2001. Keiretsu and relationship-specific investment: A barrier to trade? *International Economic Review* 42: 871–901.
Symeonidis, G. 2003. Comparing Cournot and Bertrand equilibria in a differentiated duopoly with product R&D. *International Journal of Industrial Organization* 21: 39–53.
Toshimitsu, T. 2003. Optimal R&D policy and endogenous quality choice. *International Journal of Industrial Organization* 21: 1159–1178.
Vickers, J. 1985. Delegation and the theory of the firm. *Economic Journal* 95: 138–147.
Vives, X. 1984. Duopoly information equilibrium: Cournot and Bertrand. *Journal of Economic Theory* 34: 71–94.
Vives, X. 2005. Complementarities and games: New developments. *Journal of Economic Literature* 43: 437–479.
Zhao, J. 2001. A characterization for the negative welfare effects of cost reduction in a Cournot oligopoly. *International Journal of Industrial Organization* 19: 455–469.
Ziv, A. 1993. Information-sharing in oligopoly: The truth-telling problem. *RAND Journal of Economics* 24: 455–465.

Cournot, Antoine Augustin (1801–1877)

Martin Shubik

Keywords

Bertrand, J. L. F.; Chamberlin, E. H.; Chance; Cournot, A. A.; Edgeworth cycle; Entry; Equation of exchange; Jevons, W. S.; Large

group equilibrium; Law of demand; Marshall, A.; Mathematical economics; Monopolistic competition; Monopoly; Non-cooperative equilibrium; Objective and subjective probability; Oligopolistic competition; Oligopoly; Price as a strategic variable; Probability; Product differentiation; Quantity as a strategic variable; Silver standard; Supply and demand; Uncertainty; Unlimited competition; Value in exchange; Walras, L.; Wealth

JEL Classifications
B31

Cournot was born at Gray (Haute-Saône) on 28 August 1801 and died in Paris on 30 March 1877. Until the age of 15 his education was at Gray. After studying at Besançon he was admitted to the Ecole Normale Supérieure in Paris in 1821. In 1823 he obtained his licentiate in sciences and in October of that year was employed by Marshal Gouvion-Saint-Cyr as literary adviser to the Marshal and tutor to his son. In 1829 he obtained his doctorate in science with a main thesis in mechanics and a secondary one in astronomy. Through the sponsorship of Poisson in 1834 he obtained the professorship in analysis and mechanics at Lyon.

After a year of teaching he became primarily involved in university administration. In 1835 he became rector of the Académie de Grenoble and subsequently became inspector general of education and from 1854 to 1862 was rector of the Académie de Dijon. He became a Knight of the Legion of Honour in 1838 and an Officer in 1845. He was afflicted with failing eyesight and in the last part of his life was nearly blind. In 1862 he retired from public life but continued his own researches in Paris until his death.

Cournot was a prolific writer. His writings can be broadly divided into three categories: (1) mathematics; (2) economics and (3) the philosophy of science and philosophy of history.

In considering Cournot as an economist it is necessary to place his major economic work, *Recherches sur les principes mathématiques de la théorie des richesses* (1838) in the context not only of *Principes de la théorie des richesses* (1863), which can be regarded as a literary version of his work of a quarter of a century earlier, and his *Revue sommaire des doctrines économiques* (1877) which appeared in the last year of his life, but also of his writings on probability and the philosophy of science, in particular *Exposition de la théorie des chances et des probabilités* (1843) and *Matérialisme, vitalisme, rationalisme*: *Etudes des données de la science en philosophie* (1875).

It is possible to weave a broad cloth of interpretation taking into account not merely Cournot's other works but what appears to be known of his personality and the considerable social and political flux in France during the times in which he lived. Guitton (1968) has suggested that Cournot had a rather melancholic and solitary temperament and 'did nothing to make his books attractive'. He notes that: 'Cournot was a pioneer. He did nothing to court his contemporaries, and they, in turn, not only failed to appreciate him but ignored him.' Palomba ([1981], 1984) provides a sketch of the historical background of his time, noting the growth of socialist ideas in Europe, the political actions and reactions to the French Revolution and the challenges to the concept of ownership. Rather than challenge or repeat the broad contextual interpretation of Cournot provided by Palomba, this article is confined primarily to the direct interpretation of his works in economics and supporting texts in the light of many of the developments in economics which are consistent with and may be indebted to his original ideas.

The texts followed here include the French given in the complete works of Cournot (1973) and the Nathaniel T. Bacon translation (1899) entitled *Researches into the Mathematical Principles of the Theory of Wealth,* which also contains an essay by Irving Fisher on Cournot and Mathematical Economics as well as a bibliography on Mathematical Economics from 1711 to 1897. The 1929 reprint of the 1897 edition was used.

The preface sets forth with great clarity Cournot's fundamental approach to political economy. He states:

> But the title of this work sets forth not only theoretical researches; it shows also that I intend to apply to them the forms and symbols of mathematical analysis. Most authors who have devoted themselves to political economy seem also to have had a wrong idea of the nature of the applications of mathematical analysis to the theory of wealth.
>
> But those skilled in mathematical analysis know that its object is not simply to calculate numbers, but that it is also employed to find the relations between magnitudes which cannot be expressed in numbers and between functions whose law is not capable of algebraic expression. Thus the theory of probabilities furnishes a demonstration of very important propositions, although without the help of experience it is impossible to give numerical values for contingent events, except in questions of mere curiosity, such as arise from certain games of chance. (p. 3)

Cournot continues in the preface to note that only the first principles of differential and integral calculus are required for his treatise. Professional mathematicians could be interested in it for the questions raised rather than the level of mathematics presented. He ends the preface with the caveat:

> I am far from having thought of writing in support of any system, and from joining the banners of any party; I believe that there is an immense step in passing from theory to governmental applications; I believe that theory loses none of its value in thus remaining preserved from contact with impassioned polemics; and I believe, if this essay is of any practical value, it will be chiefly in making clear how far we are from being able to solve, with full knowledge of the case, a multitude of questions which are boldly decided every day. (p. 5)

The first chapter, 'Of Value in Exchange or of Wealth in General', provides insight into the breadth of Cournot's concern for the social and historical context of wealth.

> Property, power, the distinctions between masters, servants and slaves, abundance, and poverty, rights and privileges, all these are found among the most savage tribes, and seem to flow necessarily from the natural laws which preside over aggregations of individuals and of families; but such an idea of wealth as we draw from our advanced state of civilization, and such as is necessary to give rise to a theory, can only be slowly developed as a consequence of the progress of commercial relations, and of the gradual reaction of those relations on civil institutions. (pp. 7–8)

He notes that: 'it is a long step to the abstract idea of *value in exchange* which supposes that the objects to which such value is attributed *are in commercial circulation*.'

In order to illustrate the distinction between the word *wealth* in ordinary speech and value in exchange, he presents an example of a publisher who destroys two-thirds of his stock expecting to derive more profit from the remainder than the entire edition. The economics of elasticity is developed more formally in Chapter 4 on demand, but the concept is clear.

Chapter 2, 'On Changes in Value, Absolute and Relative', begins by noting that 'we can only assign value to a commodity by reference to other commodities'. This leads to a discussion of the use of a corrected money which would serve as 'the equivalent of the mean sun of the astronomers'.

Chapter 3, 'Of the Exchanges', is the first in which formal mathematical manipulation is employed. He considers a silver standard in which all currencies are fixed in ratio to a gram of fine silver. He observes that the ratios of exchange for the same weight of fine silver cannot differ by more than transportation and smuggling costs. Given the volume of trade measured in silver he considers the arbitrage conditions for the $m(m-1)/2$ ratios among m centres. Fisher (1892) notes, however, that Cournot did not appear to be acquainted with determinants as he did not attempt a general solution of the exchange equations he proposed, but limited his calculations to three centres of exchange.

It is in Chapter 4, 'On the Law of Demand', that the modernity of his approach stands out. He is interested in demand as it is revealed in sales at a given price. He represents the relationship between sales and price by the continuous function $D = F(p)$ and observes that this function generally increases in size with a fall in price and that the empirical problem is to determine the form of $F(p)$. He indicates an appreciation of the concept of elasticity of demand although he did not develop the formal measure.

Chapters 5 and 6 deal with monopoly without and with taxation; Chapter 7 is on the competition of producers and Chapter 8 on unlimited competition. The ninth chapter is on the mutual relations of producers and the tenth on the communication

of markets. The final two chapters are somewhat macroeconomic in scope. Chapter 11 is entitled 'Of the Social Income' and 12 'Of Variations in the Social Income, Resulting from the Communication of Markets'.

As our commentary is primarily on Chapters 5–8, the order is reversed and 11 and 12 are dealt with first. Cournot explicitly avoids setting up the whole closed microeconomic system.

> It seems, therefore, as if, for a complete and rigorous solution of the problems relative to some parts of the economic system, it were indispensable to take the entire system into consideration. But this would surpass the powers of mathematical analysis and of our practical methods of calculation, even if the values of all the constants could be assigned to them numerically. The object of this chapter and of the following one is to show how far it is possible to avoid this difficulty, while maintaining a certain kind of approximation, and to carry on, by the aid of mathematical symbols, a useful analysis of the most general questions which this subject brings up.
>
> We will denote by *social income* the sum, not only of incomes properly so called, which belong to members of society in their quality of real estate owners or capitalists, but also the wages and annual profits which come to them in their capacity of workers and industrial agents. We will also include in it the annual amount of the stipends by means of which individuals or the state sustain those classes of men which economic writers have characterized as unproductive, because the product of their labour is not anything material or saleable. (pp. 127–8)

But, using a first order approximation, he studies the effect of a change in price and consumption of a good on social income as a whole under competition, under monopoly and when a new product is introduced.

> Finally, although we make continuous and almost exclusive use of the word *commodity,* it must not be lost sight of (Article 8) that in this work we assimilate to commodities the rendering of services which have for their object the satisfaction of wants or the procuring of enjoyment. Thus when we say that funds are diverted from the demand for commodity A to be applied to the demand for commodity B, it may be meant by this expression that the funds diverted from the demand for a commodity properly so called, are employed to pay for services or vice versa. When the population of a great city loses its taste for taverns and takes up that for theatrical representations, the funds which were used in the demand for alcoholic beverages go to pay actors, authors, and musicians, whose annual income, according to our definition, appears on the balance sheet of the social income, as well as the rent of the vineyard owner, the vine-dresser's wages, and the tavern-keeper's profits. (p. 149)

The last chapter considers international trade and national income and uses a first order approximation rather than a closed equilibrium system to study the benefits of opening up trade.

> Moreover (and this is the favourite argument of writers of the school of Adam Smith), it should be inferred from the asserted advantage assigned to the exporting market, and the asserted disadvantage suffered by the importing market, that a nation should so arrange as always to export and never to import, which is evidently absurd, as it can only export on condition of importing, and even the sum of the values exported, calculated at the moment of leaving the national market, must necessarily be equal to the sum of the values imported, calculated at the moment of arrival on the national market. (p. 161)

Cournot also notes the problem of analysing a tariff war:

> The question would no longer be the same if establishment of a barrier for the benefit of A producers might provoke, by way of retaliation, the establishment of another barrier for the benefit of B producers, against whom the first barrier was raised. The government of A would then have to weigh the advantage resulting from the first measure to the citizens of A against the drawbacks caused by the retaliation. The two markets A and B would thus again be placed in symmetrical conditions, and each should be considered as acting the double part of an exporting and importing market. (p. 164)

He closes his comments with:

> We have just laid a finger on the question which is at the bottom of all discussions on measures which prohibit or restrict freedom of trade. It is not enough to accurately analyse the influence of such measures on the national income; their tendency as to the distribution of the wealth of society should also be looked into. We have no intention of taking up here this delicate question, which would carry us too far away from the purely abstract discussions with which this essay has to do. If we have tried to overthrow the doctrine of Smith's school as to barriers, it was only from theoretical considerations, and not in the least to make ourselves the advocates of prohibitory and restrictive laws. Moreover, it must be recognized that such questions as that of commercial liberty are not settled either by the arguments of scientific men or even by the wisdom of statesmen. (p. 171)

He closes his work with the observation about theory that:

> By giving more light on a debated point, it soothes the passions which are aroused. Systems have their fanatics, but the science which succeeds to systems never has them. Finally, even if theories relating to social organization do not guide the doings of the day, they at least throw light on the history of accomplished facts. (p. 171)

Although the contribution of these last chapters is not as great as those to which we now turn, the spirit and style is that of a major theorist concerned deeply and objectively with application to practical affairs.

In Chapters 5–9 Cournot develops his theory of monopoly, oligopoly and unlimited competition. This can be contrasted with Ricardo (1817) before and Walras (1874) after, who concentrated on unlimited competition with no aim at producing a unified theory involving numbers.

In Chapter 5 Cournot deals with monopoly, considering increasing, decreasing and constant returns and in Chapter 6 the influence of taxation on a monopoly is considered. He notes direct taxes and indirect taxes as well as bounties and their influences on both producers and consumers; and closes with an examination of two variations of taxation in kind.

Chapter 7 provides a smooth transformation from single person maximization to non-cooperative optimization where agents who mutually influence each other act without explicit cooperation.

> We say *each independently,* and this restriction is very important, as will soon appear; for if they should come to an agreement so as to obtain for each the greatest possible income, the results would be entirely different, and would not differ, so far as consumers are concerned, from those obtained in treating of a monopoly.
>
> Instead of adopting $D = F(p)$ as before, in this case it will be convenient to adopt the inverse notation $p = f(D)$; and then the profits of proprietors (1) and (2) will be respectively expressed by
>
> $$D_1 f(D_1 + D_2), \text{and } D_2 f(D_1 + D_2),$$
>
> i.e. by functions into each of which enter two variables, D_1 and D_2. (p. 80)

It is at this point that Cournot switches from price to quantity of a homogeneous product as the strategic variable used by the competitors. His words and the mathematics do not quite match. He says, 'This he will be able to accomplish by properly adjusting his price.' The first order condition for the existence of a non-cooperative equilibrium with quantity as the strategic variable is given. A diagram showing a stable equilibrium and another with a non-stable equilibrium are presented. The analysis is generalized to n producers including the possibility of an extra group of producers beyond n, all of whom produce at capacity. He obtains n symmetric equations for the firms with interior production levels and sets the others at capacity.

When he introduces n different general cost functions for the n firms he handles the situation with all having an equilibrium defined by the simultaneous satisfaction of the equations arising from the first order conditions. But he does not deal with the possibility that costs could be such that different subsets of firms could be active in different equilibria.

The criticism levelled by Bertrand (1883) in his review written well after Cournot's death concerns the modelling rather than the mathematics. As Cournot considered competition without entry among firms selling an identical product it was fairly natural to avoid the discontinuity in the payoff function caused by selecting price as an independent variable. But the observation of Bertrand matters for markets with a finite number of firms. The choice of strategic variable causes not only mathematical difficulties but raises questions concerning economic realism and relevance. Quantity, price, quality, product differentiation and scope can all be considered as playing dominant roles in different markets. But the general explanation of price and quantity as strategic variables was and is critical to the development of economic theory. Cournot provided the foundations for the understanding of quantity. Bertrand, whose review of the books of Cournot and Walras was somewhat tangential to his professional interests offered only an example rather than a developed theory of price competition. It remained for Edgeworth (1925, pp. 111–42) to explore the underlying difficulties with the payoff functions for duopoly with increasing marginal costs; and it

has only been since the 1950s with the advent of the theory of games that there has been an adequate study of the properties of non-cooperative equilibria in games with price and quantity as strategic variables, without or with product differentiation.

The thesis of Nash (1951) on the existence of non-cooperative equilibria for a class of games in strategic form provided a broad general underpinning for the concept of non-cooperative equilibrium. It was then immediately observable that, although Cournot's work with equilibria of games with a continuum of strategies was not strictly covered by Nash's work, conceptually Cournot's solution could be viewed as an application of non-cooperative equilibrium theory to oligopoly (see Mayberry et al. 1953). The broader investigation of the price model and the interpretation of the instability of the Edgeworth cycle in terms of mixed strategy equilibria has only taken place recently. This also includes a growing literature on how to embed both the Cournot and Bertrand–Edgeworth models into a closed economic system or Walrasian framework. A summary of much of this work is presented by Shubik (1984).

It is important to appreciate that the developments in the theory of monopolistic competition such as those of Hotelling (1929) and Chamberlin (1933) and J. Robinson (1933) were based upon the Cournot non-cooperative game model. Although it may be argued that Chamberlin's and Mrs. Robinson's works possibly contained broader and richer models of competition among the few than that of Cournot, they represented a step backwards in their lack of mathematical sophistication and analysis. The Chamberlin discussion of large group equilibrium does have price as the strategic variable along with product differentiation and entry, but the solution concept is the non-cooperative equilibrium à la Cournot with the caveat that an attempt to produce a strict formal mathematical model of Chamberlin's large group equilibrium leads one to conclude that the game having price as a strategic variable is closer to Edgeworth's analysis than that of Cournot and a price strategy non-cooperative equilibrium may not exist.

In Chapter 8 Cournot shows his basic grasp of the important strategic difference between pure competition and oligopolistic competition. Using his own words, he states:

> The effects of competition have reached their limit, when each of the partial productions D_2 is *inappreciable,* not only with reference to the total production $D = F(p)$, but also with reference to the derivative $F'(p)$, so that the partial production D_k could be subtracted from D without any appreciable variation resulting in the price of the commodity. This hypothesis is the one which is realized, in social economy, for a multitude of products, and, among them, for the most important products. It introduces a great simplification into the calculations, and this chapter is meant to develop the consequences of it. (p. 90)

In modern mathematical economics, in the linking of competition among the few and the Walrasian system into a logically consistent whole, two approaches to the study of large numbers have been adopted. The first is replication and has its roots in Cournot and, more formally, Edgeworth (1881). Following Edgeworth this method was used in cooperative core theory by Shubik (1959). The second involves considering a continuum of economic agents where each agent can be regarded as a set of measure zero. Cournot clearly saw the need to consider a market in which each individual firm is too small to influence price. But it remained for Aumann (1964) to fully formalize the concept of an economic game with a continuum of agents.

After 25 years during which his seminal work in mathematical economics was essentially ignored, Cournot demonstrated his concern for his ideas by publishing *Principes de la théorie des richesses* (1863), where he offered a non-mathematical rendition of his early work. This book is of considerably greater length than its predecessor and is divided into four books: Book 1, Les Richesses (eight chapters); Book 2, Les Monnaies (seven chapters); Book 3, Le Système économique (ten chapters) and Book 4, L'Optimisme économique (seven chapters).

This book met with no more immediate success than his original work and is not as deep. For example the chapters on money, although they contain discursive and historical material of interest, have little material of analytic depth.

In spite of the indifference of the environment to his writings in economics, Cournot regarded his contribution as sufficiently important that some 14 years later, in the year of his death, he published his *Revue sommaire des doctrines économiques* (1877). This book was also longer, non-mathematical and of less significance than the work of almost 40 years earlier. But Cournot's own sense of having been at least partially vindicated after 40-odd years is indicated in his *avant-propos*:

> I was at that point in 1863, when I had the desire to find out whether I had sinned in the substance of ideas or only in their form. To that end, I went back to my work of 1838, expanding it where needed, and, most of all, removing entirely the algebraic apparatus which intimidates so much in these subjects. Whence the book entitled: 'Principes de la théorie des richesses'. 'Since it took me,' I said in the preface, 'twenty-five years to lodge an appeal of the first sentence, it goes without saying that I do not intend, whatever happens, to resort to any other means. If I lose my case a second time, I will be left only with the consolation which never abandons disgraced authors: that of thinking that the sentence that condemns them will one day be quashed in the interest of the law, that is of the truth.'
>
> When I took this engagement in 1863, I did not think that I would live long enough to see my 1838 case reviewed as a matter of course. Nevertheless, more than thirty years later, another generation of economists, to put it like Mr. the commander Boccardo, discovered that I opened up back then, though too timidly and too partially, a good path to be followed, on which I was even somewhat preceded by a man of merit, the doctor Whewell. While another Englishman, Mr. Jevons, was undertaking to enlarge this path, a young Frenchman, Mr. Leon Walras, professor of Political Economy at Lausanne, dared to maintain right in the Institute that it was wrong to pay so little attention to my method and my algorithm, which he used rightfully to expose a new theory, more amply developed.
>
> Now, look at my bad luck. If I won a little late, without any involvement, my 1838 case, I lost my 1863 case. If one wanted in retrospective to make a case for my algebra, my prose (I am ashamed of saying it) did not get better success from the publisher. The *Journal des Economistes* (August 1864) criticized me mainly 'for not having moved on from Ricardo,' for not having taken into account the discoveries that so many men of merit have made in twenty-five years in the field of political economy; thus the poor author that no one of the official world of French economists wanted to quote incurred the reproach of not having quoted others enough.

Cournot was central to the founding of modern mathematical economics. The average reader tends not to be aware that the textbook presentations of the 'marginal cost equals marginal revenue' optimizing condition for monopoly and 'marginal cost equals price' for the firm in pure competition come directly from the work of Cournot (including an investigation of the second order conditions).

He had to wait many years for recognition, but when it came in the works of Jevons, Marshall, Edgeworth, Walras and others, it moved the course of economic theory. Marshall notes (*Memorials of Alfred Marshall,* pp. 412–13, letter 2, July 1900) 'I fancy I read Cournot in 1868', this was when Marshall was 26, some 30 years after the book appeared. He acknowledges him both as a great master and as his source 'as regards the form of thought' for Marshall's theory of distribution. Jevons, in his preface to the second edition of *The Theory of Political Economy* records 'I procured a copy of the work as far back as 1872' and that it 'contains a wonderful analysis of the laws of supply and demand, and of the relations of prices, production, consumption, expenses and profits'. He excuses himself for his lateness in coming to Cournot observing: 'English economists can hardly be blamed for their ignorance of Cournot's economic works when we find French writers equally bad.' Walras in the preface to the fourth edition of *Elements of Pure Economic* (Jaffé translation, p. 37) acknowledges his 'father Auguste Walras, for the fundamental principles of my economic doctrine'; and 'Augustin Cournot for the idea of using the calculus of functions in the elaboration of this doctrine'. His liberal references to Cournot include his discussion of monopoly and the description of supply and demand.

The art of formal modelling is different from but related to the use of mathematical analysis in

economics. The clarity and parsimony of Cournot's modelling stand out and have served as beacons guiding the development of mathematical economics.

An important feature missing from Cournot's seminal work is the discussion of the role of chance and uncertainty in the economy. He stressed the importance of chance in both his book *Exposition de la théorie des chances et des probabilités* (1843) and in *Matérialisme, vitalisme, rationalisme* (1875).

Although economics was the only social science he attempted to mathematize, he was well aware of the simplifications being made in cutting economic analysis from the context of history and society.

> The economist considers the body social in a state of division and so to say of extreme pulverization, where all the particularities of organization and of individual life offset each other and vanish. The laws that he discovers or believes to discover are those of a mechanism, not those of a living organism. For him, it is no longer a question of social physiology, but of what is rightfully called social physics (p. 56). We mention that these cases of regression which imply abstractions of the same kind, if not of the same type and of the same value, reappear in various stages of scientific construction.

Cournot's work on chance and probability does not appear to have provided any new mathematical analysis, but he made three distinctions concerning the nature of probability. His book of 1843 was a text with the dual purpose of teaching the non-mathematician the rules of the calculus of probability and of dissipating the obscurities on the delicate subject of probability. He stressed the distinction between objective and subjective probability. His opening chapters provide a discussion of the appropriate combinatorics and frequency of occurrence interpretation of probability.

Cournot stressed the distinction between objective probability where frequencies are known and subjective probability. He noted:

> We could, since then, relying on the theorems of Jacques Bernoulli, who was already aware of their meaning and scope, pass immediately to the applications those theorems had in the sciences of facts and observations. However, a principle, first stated by the Englishman Bayes, and on which Condorcet, Laplace and their successors wanted to build the doctrine of 'a posteriori' probabilities, became the source of much ambiguity which must first be clarified, of serious mistakes which must be corrected and which are corrected as soon as one has in mind the fundamental distinction between probabilities which have an objective existence, which give a measure of the possibility of things, and subjective probabilities, relating partly to one's knowledge, partly to one's ignorance, depending on one's intelligence level and on the available data. (p. 155)

Subjective probability rests on the consideration of events which our ignorance calls for us to treat as equiprobable due to insufficient cause.

He added a third category which he entitled 'philosophical probability' (Chapter 17) 'where probabilities are not reducible to an enumeration of chances' but 'which depend mainly on the idea that we have of the simplicity of the laws of nature' (p. 440).

Cournot's views on probability appear to be intimately related to his concern for social statistics and economic modelling. Although he did not establish formal links between his mathematical economics models and chance he regarded history and the development of institutions as dependent on chance and economics as set in the context of institutions.

Cournot was at best an indifferent mathematician. Bertrand clearly dominated him in that profession. But from his own writings it is clear that Cournot was well aware of both his purpose in applying mathematics to economics and his limitations as a mathematician. At the age of 58 he wrote his *Souvenirs* which he finished in Dijon in October 1859. They were published many years later with an introduction by Botinelli (1913). In these writings Cournot provides his self-assessment as a mathematician.

> I was starting to be a little known in the academic world through a fairly large number of scientific articles. This was the basis of my fortune. Some of these articles ended up with Mr. Poisson, who was then the leader in Mathematics at the Institute, and mainly at the University, and he liked them

particularly. He found in them philosophical insight, which I think was not all that wrong. Furthermore, he foresaw that I would go a long way in the field of pure mathematical speculation, which was (I always thought it and never hesitated to say it) one of his mistakes.

The general tenor of his *Souvenirs* is of a moderately conservative, quietly humourous, self-effacing man with considerable understanding of his environment and a broad belief in science and its value to society.

Regarding his work as a whole, his dedication and power as the founder of mathematical economics and the promoter of empirical numerical investigations emerges. He strove for around 40 years to have his ideas accepted. He did so with persistence and humour (referring to his major work as 'mon opuscule'). He understood the need to wait for a generation to die. And before his death with the work and words of Jevons and Walras he saw the vindication of his approach.

See Also

► Bertrand, Joseph Louis François (1822–1900)

Selected Works

1838. *Researches into the mathematical principles of the theory of wealth*. Trans. N.T. Bacon. New York: Macmillan, 1929.
1841. *Traité élémentaire de la théorie des fonctions et du calcul infinitésimal*. 2nd ed. Paris: Hachette, 1857.
1843. *Exposition de la théorie des chances et des probabilités*. Paris: Hachette.
1861. *Traité de l'enchaînement des idées fondamentales dans les sciences et dans l'histoire*. New ed. Paris: Hachette, 1911.
1863. *Principes de la théorie des richesses*. Paris: Hachette.
1872. *Considérations sur la marche des idées et des évènements dans les temps modernes*. 2 vols. Paris: Boivin, 1934.
1875. *Matérialisme, vitalisme, rationalisme: Études des données de la science en philosophie*. Paris: Hachette, 1923.
1877. *Revue sommaire des doctrines économiques*. Paris: Hachette.
1913. *Souvenirs 1760–1860.* With an introduction by E.P. Bottinelli. Paris: Hachette. Published posthumously.
1973. *A.A. Cournot Oeuvres Complètes,* 5 vols, ed. André Robinet. Paris: Librairie Philosophique J. Vrin.

Bibliography

Aumann, R.J. 1964. Markets with a continuum of traders. *Econometrica* 32: 39–50.
Bertrand, J.L.F. 1883. (Book reviews of) *Théories Mathematique de la richesse sociale* par Léon Walras; *Recherches sur les principes mathématiques de la théorie de la richesse* par Augustin Cournot. *Journal des Savants* 67: 499–508.
Chamberlin, E.H. 1933. *The theory of monopolistic competition*. Cambridge, MA: Harvard University Press.
Edgeworth, F.Y. 1881. *Mathematical psychics*. London: Kegan Paul.
Edgeworth, F.Y. 1925. *Papers relating to political economy, I*. London: Macmillan.
Fisher, I. 1892. *Mathematical investigations in the theory of value and prices*. New Haven: Connecticut Academy of Arts and Sciences, Transactions 9. Reprinted, New York: Augustin M. Kerlley, 1961.
Guillebaud, C.W., ed. 1961. *Marshall's principles of economics*, Notes. Vol. 2. London: Macmillan.
Guitton, H. 1968. Antoine Augustin Cournot. In *The international encyclopedia of the social sciences,* vol. 3. New York: Macmillan and Free Press.
Hotelling, H. 1929. Stability in competition. *Economic Journal* 34: 41–57.
Jevons, W.S. 1911. *The theory of political economy.* 4th ed, 1931. London: Macmillan.
Mayberry, J., J.F. Nash, and M. Shubik. 1953. A comparison of treatments of a duopoly situation. *Econometrica* 21: 141–155.
Nash, J.F. Jr. 1951. Noncooperative games. *Annals of Mathematics* 54: 289–295.
Palomba, G. 1984. Introduction à l'oeuvre de Cournot. *Economie Appliquée* 37: 7–97. Trans. from Italian, extracted from *Cournot Opere*, Turin: UTET (1981).
Ricardo, D. 1817. *The principles of political economy and taxation*, 1965. London: J.M. Dent.
Robinson, J. 1933. *The economics of imperfect competition*. London: Macmillan.
Shubik, M. 1959. Edgeworth market games. In *Contributions to the theory of games IV*, ed. A.W. Tucker and R.D. Luce. Princeton: Princeton University Press.
Shubik, M. 1984. *A game theoretic approach to political economy*. Cambridge, MA: MIT Press.
Walras, L. 1874–1877. *Elements of pure economics*. Trans. W. Jaffée. London: George Allen & Unwin, 1954.

Court, Louis Mehel (Born 1910)

M. Ali Khan

Court obtained his PhD degree in economics from Columbia University in 1942. He was at Columbia from the Summer Term of 1936 through the Spring Term of 1938 and held the Granville W. Garth Fellow-ship. He was a 'student at large' at the University of Chicago during the Winter and Spring Quarters of 1941. It is also known that he was an Instructor of Mathematics at Rutgers University during the years 1946–8. This sparse and sketchy information allows some gleaning into the intellectual influences on Louis Court.

Court's first published paper appeared in the *Journal of Mathematics and Physics* and concerned what would now fall under the heading of duality in mathematical programming. In particular he showed that:

> if $\phi\ (q_{1,\ldots,}\ q_n)$ is maximized subject to the single constraint $\sum_{i=1}^{n} p_i q_i = m$ and $q_i(\ p_1\ ,\ldots,\ p_n\ )(i = 1,\ldots,\ n)$ [satisfy] the first order conditions for this maximization, then the function $\psi(p_1,\ldots,p_n)\equiv\phi[q_1(p_1,\ldots,p_n),\ \ldots,\ q_n(p_1,\ldots,p_n)]$ is minimized subject to the same constraint except that the p's are now regarded as the active variables.

Court returned to this result in his 1951 paper which he presented at the International Congress of Mathematicians and in which he not only generalized his result but also noted its relevance to isoperimetric problems in the calculus of variations and to possible applications concerning integrability problems in the theory of differential equations. He had by then already applied his result to statistical decision, theory. One can only wonder why he did not pursue applications in economic theory.

Louis Court's place among the pioneers of mathematical economics thus rests on his 1941 *Econometrica* articles in which he extended the theories of consumer and producer behaviour to a setting with infinitely many commodities. He introduced his paper with the following rather modern statement:

> Apart from its utility in treating commodity groups embracing large, though not necessarily infinite numbers of items, the extension is stamped with true intellectual concinnity. The finite theories are contained, as very special cases, in the infinite analyses. . . . Housing provides an instance in which it is profitable to use the commodity-spectrum concept.

Court worked in the space of square Lebesgue integrable functions over a compact interval; saw the relevance of the theory of Hilbert spaces, still in its infancy; formulated the price system as an element of the topological dual of his commodity space and emphasized the distinction between functions and their equivalence classes (see ibid., footnote 5, p. 248). However, he did not see the relevance of Ramsey's (1928) work or the importance of the weak, weak star and Mackey topologies, the latter omission being justified by the fact that the basic papers dealing with these concepts had yet to be written. Court's results were extended to reflexive Banach spaces by Berger (1971).

In summary, Court's *Econometrica* papers may be seen as a first serious application of functional analysis (see, for example, Dieudonné 1981) to economic theory and as a precursor of Debreu (1954), Hurwicz (1958), Bewley (1972) and of the burgeoning literature inspired by these contributions; and in another context, of Dornbush–Fischer–Samuelson (1977). The evaluation of the (then) editor of *Econometrica*, Ragnar Frisch, still stands:

> Even though considerable portions of [the] mathematical technique are in essence the same as that developed by Volterra and others, a presentation of this technique shaped especially with the econometric problems in view, is highly useful. Economic theory is now growing into a stage where much of the work will consist of a combination of mathematical and economic analyses so intimate that it is difficult to say where one begins and the other ends. Mr. Court's paper is a valuable contribution towards this type of work.

See Also

► Functional Analysis

Selected Works

1941. A theorem on maxima and minima with an application to differential equations. *Journal of Mathematics and Physics* 20: 99–106. Reviewed by J. Reid in *Mathematical Reviews* 2: 287.
1941. Entrepreneurial and consumer demand theories for commodity spectra. *Econometrica* 9: 135–162 and 241–297.
1944. A reciprocity principle for the Neyman–Pearson theory of testing statistical hypotheses *Annals of Mathematical Statistics* 15: 326–327. Reviewed by J. Wolfowitz in *Mathematical Reviews* 6: 93.
1951. A theorem on conditional extremes with an application to total differentials. Proceedings at the *American Mathematical Society* 2: 423–428. Reviewed by J. Reid in *Mathematical Reviews* 13: 215.

Bibliography

Berger, M.S. 1971. Generalized differentiation and utility functionals for commodity spaces of arbitrary dimension. In *New York: Preferences, utility, and demand*, ed. J.S. Chipman et al. New York: Harcourt Brace Jovanovich.
Bewley, T.F. 1972. Existence of equilibria in economies with infinitely many commodities. *Journal of Economic Theory* 4: 514–540.
Dieudonné, J. 1981. *History of functional analysis*. Amsterdam: North-Holland.
Dornbusch, R., S. Fischer, and P. Samulson. 1977. Comparative advantage, trade and payments in a Ricardian model with a continuum of goods. *American Economic Review* 67: 823–839.
Hurwicz, L. 1958. Programming on linear spaces. In *Studies in linear and non-linear programming*, ed. K.J. Arrow et al. Stanford: Stanford University Press.
Ramsey, F.P. 1928. A mathematical theory of saving. *Economic Journal* 38: 543–559.

Covered Interest Parity

C. Emre Alper and Oya Pinar Ardic

Abstract

Covered Interest Parity describes an idealised situation in foreign exchange markets in which the interest rates on assets differing only in the currency of denomination will be equal. This article describes the theoretical assumptions under which CIP holds and the evidence for CIP in practice.

Keywords

CIP; Counterparty risk; Emerging markets; Foreign exchange

JEL Classifications

F31

The Covered Interest Parity (CIP) condition states that the actions of foreign exchange market participants, when hedged against exchange rate risk using the forward exchange market, should equalise interest rates on any two assets that differ only in currency of denomination. The assumptions under which the CIP condition holds are (i) negligible transaction costs, (ii) perfect capital mobility, (iii) many participants in the spot and the forward exchange markets with ample funds and no counterparty risks, and (iv) identical default and political riskiness, liquidity, maturity and seniority of the underlying assets. These assumptions in essence rule out transactions involving smaller currencies, the existence of capital controls or taxes on financial flows, as well as thin markets and periods of crisis.

Algebraically, suppose that $i_{t,k}$ and $i^*_{t,k}$ denote the interest rates on domestic currency and foreign-currency-denominated assets at time t, respectively, for an investment horizon of k. Suppose also that S_t denotes the spot exchange rate at time t, i.e. the current price of one unit of foreign currency in domestic currency units, while $F_{t,k}$ is the k-period forward exchange rate at time t, i.e. the exchange rate currently agreed upon for a transaction k periods ahead. The CIP condition states that

$$\left(1 + i_{t,k}\right) = \left(1 + i^*_{t,k}\right)\frac{F_{t,k}}{S_t} \qquad (1)$$

In most settings, the log approximation of (1) is used

$$f_t^k - s_t = i_t - i_t^* \tag{2}$$

where f and s denote the natural logarithms of F and S, respectively.

There is a vast literature that studies whether the CIP condition holds for a variety of currency pairs, asset types etc. Officer and Willett (1970) and Taylor (1992) provide essential surveys of this literature. There are two broad reasons to be interested in the validity of the CIP condition. First, it is taken as an indicator of market efficiency and international market integration. Second, many models in international finance and open economy macroeconomics assume the CIP condition and use it as a key building block. Lack of empirical support for such models might in fact be due to the failure of the CIP condition (among other possible factors).

Empirical tests of the CIP condition mainly consider larger currencies and in general involve US dollars on one side of the transaction. These tests can be broadly grouped into two based on their methodologies. The first set of studies calculate deviations from the CIP condition and analyse the time series properties of these deviations. The second conducts regression analyses of an estimable CIP relation of the form

$$f_t^k - s_t = \mathrm{A} + \beta\left(i_t - i_t^*\right) + \varepsilon_t \tag{3}$$

to test if $\alpha = 0$ and $\beta = 1$ while simultaneously testing if ε_t is independently and identically distributed.

Overall, empirical evidence from developed economies supports the validity of the CIP condition. See, for example, Frenkel and Levich (1975, 1977), and Taylor (1987, 1989) among others. Progressive financial liberalisation and the integration of international financial markets since the 1990s are likely to result in even more favourable evidence for the CIP condition. One may summarise the key observations of this literature as follows. First, deviations from the CIP condition, only to the extent that they exceed transaction costs, can be deemed significant. Second, during times of financial turbulence, when counterparty risk is especially prevalent, deviations from the CIP condition tend to be significant and persistent. Third, as opposed to monthly averages or daily closing values, real-time data should be used, as this is what the market participants face at the time of transactions. Fourth, it is not possible to test the validity of the CIP condition over long horizons since the longest maturity forward contract publicly traded is in general for 12 months.

As for the emerging market economies, it is not yet possible to state whether the CIP condition holds or not because the four underlying assumptions of the CIP condition do not hold in general. Nevertheless, Kumhof (2001) reports, for a number of emerging markets, that although there are substantial deviations from the CIP condition in the short run, a stable relationship exists between the forward premium and the interest differential in the long run.

See Also

- ▶ Foreign Exchange Markets, History of
- ▶ Optimality and Efficiency

Bibliography

Frenkel, J.A., and R.M. Levich. 1975. Covered interest arbitrage: Unexploited profits? *Journal of Political Economy* 83: 325–38.

Frenkel, J.A., and R.M. Levich. 1977. Transaction costs and interest arbitrage: Tranquil versus turbulent periods. *Journal of Political Economy* 85: 1209–26.

Kumhof, M. 2001. International capital mobility in emerging markets: New evidence from daily data. *Review of International Economics* 9: 626–40.

Officer, L.H., and T.D. Willett. 1970. The covered interest arbitrage schedule: A critical survey of recent developments. *Journal of Money, Credit, and Banking* 2: 247–57.

Taylor, M.P. 1987. Covered interest parity: A high-frequency, high-quality data study. *Eco-nomica* 54: 429–38.

Taylor, M.P. 1989. Covered interest arbitrage and market turbulence. *Economic Journal* 99: 376–91.

Taylor, M.P. 1992. Covered interest parity. In *The new palgrave dictionary of money and finance*, ed. M. Milgate, P. Newman, and J. Eatwell, 509–11. London: Macmillan.

Crawling Peg

David Vines

A 'crawling peg' denotes an exchange rate system in which the value of a country's currency is fixed but moveable. The country would undertake to keep its currency at a fixed, or 'par' value. But that par value itself would be *gradually* changed, if this were necessary to correct a 'fundamental disequilibrium' in the country's balance of payments. As elaborated by Williamson (1965) the rate of gradual adjustment would be limited to a maximum rate of one twenty-sixth of one per cent per week. Such a proposal had earlier been put forward by Meade (1964), and the idea originally came from Harrod (see Harrod 1969, p. 92).

The reason for giving this proposal the label of 'crawling peg' should be apparent. The 'adjustable peg' of the Bretton Woods system was one in which changes in par values of exchange rates were carried out infrequently, suddenly, and in a sizeable, discrete step.

The 'crawling peg' was proposed as a system under which such par changes as occur are implemented slowly, in such a large number of small steps to make the process of exchange rate adjustment continuous for all practical purposes; a system therefore under which the peg crawls from one level to another (Williamson 1965, p. 2).

If a 'crawling peg' system were not to give rise to large, and possibly disorderly, international capital flows, it would need to be accompanied by an appropriate interest rate policy. For example, if a country's exchange rate were crawling downwards by two per cent per year, then its interest rate would need to be two per cent higher than in other countries whose exchange rates were not moving, in order to avoid stimulating capital outflow.

The 'crawling peg' offered, in the late 1960s, considerable attractions. The 'adjustable peg' regime of the Bretton Woods system was then beginning to disintegrate under the influence of speculative capital flows. The trouble with the adjustable-peg system was that it delayed exchange rate adjustment until the point at which it had become a near certainty, encouraging a speculative attack which then precipitated the inevitable crisis (and which handed speculators a one way bet on a substantial capital gain). Such speculative attacks in the end broke the Bretton Woods system, and ushered in the era of floating exchange rates, with all its difficulties. The 'crawling peg' would have allowed countries to defend par values for their currencies and yet change these par values themselves without disrupting the whole system. A number of countries have, in fact, used the 'crawling peg' at some time (including Argentina, Brazil, Chile, Columbia, Israel, Uruguay and Vietnam). And there is 'rather general agreement' that it has succeeded in allowing them to neutralize efficiently the effects on their balance of payments of high inflation rates (see Williamson 1977). Although these countries are underdeveloped, their favourable experience with the crawling peg may be relevant to the major industrialized nations.

The great difficulty about a 'crawling peg' system is, however, that it makes it very difficult, or even impossible, to use interest rate policy in the pursuit of domestic economic management. Consider a country in balance of payments deficit which was also experiencing the threat of unemployment. The exchange rate would crawl downwards (because of the deficit) and this would require a relatively high level of domestic interest rates (in order to prevent capital outflow) which would be inconsistent with combatting unemployment. If, instead, interest rates were lowered to combat the unemployment, then capital would flow out. In that case either the rate of downward crawl would increase (and become a cumulative downward spiral) or the capital outflow would become a torrent (as speculators anticipated the defeat of the crawling peg system by means of a large instant currency collapse). An implied great disadvantage of the 'crawling peg' is the fact that where large changes in the exchange rate prove

necessary it is impossible to effect them immediately: this in effect means that the interest rate in the country may have to be tied down to offsetting the anticipated exchange rate change during a *very* lengthy adjustment period.

Nevertheless, the 'crawling peg' idea is still of contemporary relevance. There have been a number of recent proposals to reform the international monetary system in the direction of greater management of exchange rates, so as to limit the misalignments which are intrinsic to the present non-system of floating exchange rates. Such a new system would require 'target zones', or 'central rates', for its implementation (see Williamson 1983; Meade 1984). These target zones, or central rates, should crawl, rather than being rigidly pegged and discretely adjusted, for exactly the reasons discussed above. But the implications that this would have for domestic interest rate policy, so as to avoid destabilizing capital flows, should be clearly noted.

See Also

- ▶ Fixed Exchange Rates
- ▶ Flexible Exchange Rates
- ▶ International Finance

Bibliography

Harrod, R.F. 1969. *Money.* London: Macmillan.

Katz, S.I. 1970. The interest-rate constraint and the crawling peg. In *Exchange-rate systems, interest rates, and capital flows*, Essays in international finance, vol. 70, ed. T.D. Willet, S.I. Katz, and W.H. Branson. Princeton: Princeton University, Department of Economics, International Finance Section.

Meade, J.E. 1964. The international monetary mechanism. *Three Banks Review* 63: 3.

Meade, J. E. 1984. A Neo-Keynesian bretton woods. *Three Banks Review.*

Williamson, J.H. 1965. *The crawling peg*, Essays in international finance, vol. 50. Princeton: Princeton University, Department of Economics, International Finance Section.

Williamson, J.H. 1977. *The failure of world monetary reform, 1971–74*. Nelson: Sudbury-upon-Thames.

Williamson, J.H. 1983. *The exchange rate system*, Policy analyses in international economics, vol. 5. Washington, DC: Institute for International Economics.

Creative Destruction

Ricardo J. Caballero

C

Abstract

Creative destruction refers to the incessant product and process innovation mechanism by which new production units replace outdated ones. This restructuring process permeates major aspects of macroeconomic performance, not only long-run growth but also economic fluctuations, structural adjustment and the functioning of factor markets. Over the long run, the process of creative destruction accounts for over 50 per cent of productivity growth. At business cycle frequency, restructuring typically declines during recessions, and this add a significant cost to downturns. Obstacles to the process of creative destruction can have severe short- and long-run macroeconomic consequences.

Keywords

Banking reform; Business cycles; Canada–US Free Trade Agreement; Creative destruction; Deregulation of product markets; Economic growth; Factor reallocation; Hayek, F.; Innovation; International competition; Job flows; Job security; Labour market regulation; Liquidationist thesis; Productivity growth; Recessions; Robbins, L.; Schumpeter, J.; Structural change

JEL Classifications

D4, D10

Creative destruction refers to the incessant product and process innovation mechanism by which new production units replace outdated ones. It was coined by Joseph Schumpeter (1942), who considered it 'the essential fact about capitalism'.

The process of Schumpeterian creative destruction (restructuring) permeates major aspects of

macroeconomic performance, not only long-run growth but also economic fluctuations, structural adjustment and the functioning of factor markets.

At the microeconomic level, restructuring is characterized by countless decisions to create and destroy production arrangements. These decisions are often complex, involving multiple parties as well as strategic and technological considerations. The efficiency of those decisions not only depends on managerial talent but also hinges on the existence of sound institutions that provide a proper transactional framework. Failure along this dimension can have severe macroeconomic consequences once it interacts with the process of creative destruction (see Caballero and Hammour 1994, 1996a, b, c, 1998a, b, 2005). Some of these limitations are natural, as they derive from the sheer complexity of these transactions. Others are man-made, with their origins ranging from ill-conceived economic ideas to the achievement of higher human goals, such as the inalienability of human capital. In moderate amounts, these institutional limitations give rise to business cycle patterns such as those observed in the most developed and flexible economies. They can help explain perennial macroeconomic issues such as the cyclical behaviour of unemployment, investment and wages. In higher doses, by limiting the economy's ability to tap new technological opportunities and adapt to a changing environment, institutional failure can result in dysfunctional factor markets, resource misallocation, economic stagnation, and exposure to deep crises.

Given the nature of this short piece, I will skip any discussion of models, and refer the reader to Caballero (2006) for a review of the models behind the previous paragraph, and to Aghion and Howitt (1998) for an exhaustive survey of Schumpeterian growth models. Instead, *I* focus on reviewing recent empirical evidence on different aspects of the process of creative destruction.

Recent Evidence on the Pace of Creative Destruction

There is abundant recent empirical evidence supporting the Schumpeterian view that the process of creative destruction is a major phenomenon at the core of economic growth in market economies.

The most commonly used empirical proxies for the intensity of the process of creative destruction are those of factor reallocation and, in particular, job flows. Davis et al. (1996) (henceforth DHS) offered the clearest peek into this process by documenting and characterizing the large magnitude of job flows within US manufacturing. They defined job creation (destruction) as the positive (negative) net employment change at the establishment level from one period to the next. Using these definitions, they concluded that over ten per cent of the jobs that exist at any point in time did not exist a year before or will not exist a year later. That is, over ten per cent of existing jobs are destroyed each year and about the same amount is created within the same year. Following the work by DHS for the United States, many authors have constructed more or less comparable measures of job flows for a variety of countries and episodes. Although there are important differences across them, there are some common findings. In particular, job creation and destruction flows are large, ongoing and persistent. Moreover, most job flows take place within rather than between narrowly defined sectors of the economy.

Given the magnitude of these flows and that they take place mostly within narrowly defined sectors, the presumption is strong that they are an integral part of the process by which an economy upgrades its technology. Foster et al. (2001) provide empirical support for this presumption. They decompose changes in industry-level productivity into within-plant and reallocation (between-plant) components, and conclude that the latter – the most closely related to the creative destruction component – accounts for over 50 per cent of the ten-year productivity growth in the US manufacturing sector between 1977 and 1987. Moreover, in further decompositions they document that entry and exit account for half of this contribution: exiting plants have lower productivity than continuing plants. New plants, on the other hand, experience a learning and selection period through which they gradually catch up with incumbents. Other studies of US manufacturing based on somewhat different methodologies

(see Baily et al. 1992; Bartelsman and Dhrymes 1994) concur with the conclusion that reallocation accounts for a major component of within-industry productivity growth. Bartelsman et al. (2004) provide further evidence along these lines for a sample of 24 countries and two-digit industries over the 1990s.

Recent Evidence on the Cyclical Features of Creative Destruction

At the business cycle frequency, sharp liquidations (rises in job destruction) constitute the most noted impact of contractions on creative destruction. In contrast, job creation is substantially less volatile and mildly pro-cyclical. There is an extensive literature that, extrapolating from the spikes in liquidations (recently measured in job flows but long noticed in other contexts), finds that recessions are times of increased reallocation. In fact, this has been a source of controversy among economists at least since the pre-Keynesian 'liquidationist' theses of such economists as Hayek, Schumpeter, and Robbins. These economists saw in the process of liquidation and reallocation of factors of production the main function of recessions. In the words of Schumpeter (1934, p. 16): 'depressions are not simply evils, which we might attempt to suppress, but… forms of something which has to be done, namely, adjustment to … change.'

In Caballero and Hammour (2005) we turned the liquidationist view upside down. While we sided with Schumpeter and others on the view that increasing the pace of restructuring of the economy is likely to be beneficial, we provided evidence that, contrary to conventional wisdom, restructuring *falls* rather than rises during contractions.

Since the rise in liquidations during recessions is not accompanied by a contemporaneous increase in creation, implicit in the increased-reallocation view is the idea that increased destruction is followed by a surge in creation during the recovery phase of the cyclical downturn. This presumption is the only possible outcome in a representative firm economy, as the representative firm must replace each job it destroys during a recession by creating a new job during the ensuing recovery. However, once one considers a heterogeneous productive structure that experiences ongoing creative destruction, other scenarios are possible. The cumulative effect of a recession on overall restructuring may be positive, zero, or even negative, depending not only on how the economy contracts but also on how it recovers. Thus, the relation between recessions and economic restructuring requires one to examine the effect of a recession on aggregate separations not only at impact, but cumulatively throughout the recession-recovery episode. We explored this issue using quarterly US manufacturing gross job flows and employment data for the 1972–93 period, and found that, along the recovery path, job destruction declines and falls below average for a significant amount of time, more than offsetting its initial peak. On the other hand, job creation recovers, but it does not exceed its average level by any significant extent to offset its initial decline. As a result, our evidence indicates that, on average, recessions *depress* restructuring.

Similarly, in Caballero and Hammour (2001) we approached the question of the pace of restructuring over the cycle from the perspective of corporate assets. Studying the aggregate patterns of merger and acquisition (M&A) activity and its institutional underpinnings, we reached a conclusion that also amounts to a rejection of the liquidationist perspective. Essentially, a liquidationist perspective in this context would consider fire sales during sharp liquidity contractions as the occasion for intense restructuring of corporate assets. The evidence points, on the contrary, to briskly expansionary periods characterized by high stock market valuations and abundant liquidity as the occasion for intense M&A activity.

Recent Evidence on Institutional Impediments to Creative Destruction and Their Cost

For all practical purposes, some product or process innovation is taking place at every instant in

time. Absent obstacles to adjustment, continuous innovation would entail infinite rates of restructuring. What are these obstacles to adjustment? The bulk of it is technological – adjustment consumes resources – but (over-?) regulation and other man-made institutional impediments are also a source of depressed restructuring.

While few economists would object to the hypothesis that labour market regulation hinders the process of creative destruction, its empirical support is limited. In Caballero et al. (2004) we revisited this hypothesis using a sectoral panel for 60 countries. We found that job security provisions – measured by variables such as grounds for dismissal protection, protection regarding dismissal procedures, notice and severance payments, and protection of employment in the constitution – hamper the creative destruction process, especially in countries where regulations are likely to be enforced. Moving from the 20th to the 80th percentile in job security cuts the annual speed of adjustment to shocks by a third. By impairing worker movements from less to more productive units, effective labour protection reduces aggregate output and slows down economic growth. We estimated that moving from the 20th to the 80th percentile of job security lowers annual productivity growth by as much as 1.7 per cent.

Similarly, the idea that well-functioning financial institutions and markets are important factors behind economic growth is an old one. The process of creative destruction is likely to be a chief factor behind this link. In Caballero et al. (2006) we analysed the decade-long Japanese slowdown of the 1990s and early 2000s. The starting point of our analysis is the well-known observation that many large Japanese banks would have been out of business had regulators forced them to recognize all their loan losses. Because of this, the banks kept many zombie firms alive by rolling over loans that they knew would not be collected (evergreening). Thus, the normal competitive outcome whereby the zombies would shed workers and lose market share was thwarted. Using an extensive data-set, we documented that roughly 30 per cent of firms were on life support from the banks in 2002 and about 15 per cent of assets resided in these firms. The main idea in our article is that the counterpart to the congestion created by the zombies is a reduction in profits for potential and more productive entrants, which discourages their entry. We found clear evidence of such a pattern in firm-level data and of the corresponding reduced restructuring in sectoral data.

Bertrand et al. (2004) further drive home the point that problems in the banking sector can have grave consequences for the health of the restructuring process. They use a differences-in-differences approach on firm-level data for the period 1977–99 to analyse the impact of the banking reforms of the mid-1980s on firm and bank behaviour. These reforms eliminated government interference in bank lending decisions, eliminated subsidized bank loans, and allowed French banks to compete more freely in the credit market. They find that, after the reforms, firms' exit rates and asset reallocation rise, and are more correlated with performances.

International competition is an important source of creative destruction. Trefler (2004) concludes that there are significant productivity and reallocation effects from trade openness, even in industrialized economies. To reach this conclusion, Trefler takes advantage of the Canada–US Free Trade Agreement (FTA) to study the effects of a reciprocal trade agreement on Canada. He finds that, for industries that experienced the deepest Canadian tariff reductions, the contraction of low-productivity plants reduced employment by 12 per cent while raising industry-level labour productivity by 15 per cent. Moreover, he finds that at least half of this increase is related to exit and/or contraction of low-productivity plants. Finally, for industries that experienced the largest US tariff reductions, plant-level labour productivity soared by 14 per cent. Consistent with this evidence, Bernard et al. (2006) find that in the United States productivity growth is fastest in industries where trade costs (barriers) have declined the most.

Domestic deregulation of goods markets can have similar effects. For example, Olley and Pakes (1996) find that deregulation in the US telecommunications industry increased productivity predominantly through factor reallocation towards more productive plants rather than

through intra-plant productivity gains. More broadly, Klapper et al. (2004) study the effect of entry regulation on firm behaviour in a sample including firm-level data from countries of western and eastern Europe. Their findings support the notion that regulation affects entry: 'naturally high-entry' industries have relatively lower entry in countries that have higher entry regulations. Moreover, both the growth rate and share of high-entry industries are depressed in countries with more stringent barriers to entry. Finally, Fishman and Sarria-Allende (2004) extend the Klapper, Laeven and Rajan study to countries outside Europe and include both industry- and firm-level data from the UNIDO and WorldScope databases, and reach similar conclusions.

Final Remarks

Evidence and models coincide in their conclusion that the process of creative destruction is an integral part of economic growth and fluctuations. Obstacles to this process can have severe short- and long-run macroeconomic consequences.

See Also

- ▶ Schumpeter, Joseph Alois (1883–1950)
- ▶ Structural Change

Bibliography

Aghion, P., and P. Howitt. 1998. *Endogenous growth theory*. Cambridge, MA: MIT Press.

Baily, N., C. Hulten, and D. Campbell. 1992. Productivity dynamics in manufacturing establishments. In *Brookings papers on economic activity: Microeconomics*, ed. M. Baily and C. Winston. Washington, DC: Brookings Institution.

Bartelsman, E., and P. Dhrymes. 1994. *Productivity dynamics: US manufacturing plants, 1972–1986*, Finance and economics discussion series. Vol. 94-1. Washington, DC: Board of Governors, Federal Reserve System.

Bartelsman, E., J. Haltiwanger, and S. Scarpetta. 2004. *Microeconomic evidence of creative destruction in industrial and developing countries*. Mimeo: University of Maryland.

Bernard, A., J. Jensen, and P. Schott. 2006. Survival of the best fit: Exposure to low- wage countries and the (uneven) growth of US manufacturing plants. *Journal of International Economics* 68: 219–237.

Bertrand, M., A. Schoar, and D. Thesmar.2004. *Banking deregulation and industry structure: Evidence from the French banking reforms of 1985*. Discussion paper no. 4488. London: Centre for Economic Policy Research.

Caballero, R. 2006. *Specificity and the macroeconomics of restructuring. Yrjo Jahnsson lectures*. Cambridge, MA: MIT Press.

Caballero, R., and M. Hammour. 1994. The cleansing effect of recessions. *American Economic Review* 84: 1350–1368.

Caballero, R., and M. Hammour. 1996a. The fundamental transformation in macroeconomics. *American Economic Review* 86 (2): 181–186.

Caballero, R., and M. Hammour. 1996b. On the timing and efficiency of creative destruction. *Quarterly Journal of Economics* 111: 805–852.

Caballero, R., and M. Hammour. 1996c. On the ills of adjustment. *Journal of Development Economics* 51: 161–192.

Caballero, R., and M. Hammour. 1998a. The macroeconomics of specificity. *Journal of Political Economy* 106: 724–767.

Caballero, R., and M. Hammour. 1998b. Jobless growth: Appropriability, factor substitution and unemployment. *Carnegie-Rochester Conference Series on Public Policy* 48: 51–94.

Caballero, R., and M. Hammour. 2001. Institutions, restructuring, and macroeconomic performance. In *Advances in macroeconomic theory*, ed. J. Dreze. New York: Palgrave MacMillan.

Caballero, R., and M. Hammour. 2005. The cost of recessions revisited: A reverseliquidationist view. *Review of Economic Studies* 72: 313–341.

Caballero, R., K. Cowan, E. Engel, and A. Micco. 2004. *Effective labor regulation and microeconomic flexibility*. Mimeo, MIT.

Caballero, R., T. Hoshi, and A. Kashyap. 2006. *Zombie lending and depressed restructuring in Japan*. Working paper no. 12129. Cambridge, MA: NBER.

Davis, S., J. Haltiwanger, and S. Schuh. 1996. *Job creation and destruction*. Cambridge, MA: MIT Press.

Fishman, R., and V. Sarria-Allende. 2004. *Regulation of entry and the distortion of industrial organization*. Working paper no. 10929. Cambridge, MA: NBER.

Foster, L., J. Haltiwanger, and C. Krizan. 2001. Aggregate productivity growth: Lessons from microeconomic evidence. In *New developments in productivity analysis*, ed. E. Dean, M. Harper, and C. Hulten. Chicago: University of Chicago Press.

Klapper, L., L. Laeven, and R. Rajan. 2004. *Business environment and firm entry: Evidence from international data*. Working paper no. 10380. Cambridge, MA: NBER.

C

Olley, S., and A. Pakes. 1996. The dynamics of productivity in the telecommunications equipment industry. *Econometrica* 64: 1263–1298.
Schumpeter, J. 1934. Depressions. In *Economics of the recovery program*, ed. D. Brown et al. New York: McGraw-Hill.
Schumpeter, J. 1942. *Capitalism, socialism, and democracy*. New York: Harper & Bros.
Trefler, D. 2004. The long and short of the Canada–US Free Trade Agreement. *American Economic Review* 94: 870–895.

Creative Destruction (Schumpeterian Conception)

Arnold Heertje
Organization Name, City, UK

Keywords
Creative destruction; Dynamic efficiency; Information technology; Innovation; Restrictive practices; Schumpeter, J. A.

JEL Classifications
D4

Schumpeter invented the phrase 'creative destruction' in his famous book on the development of capitalism into socialism (Schumpeter 1942). In his view the process of creative destruction is the essential fact about capitalism and refers to the incessant mutation of the economic structure from within, destroying the old and creating a new.

In the footsteps of Karl Marx, Schumpeter argues that in dealing with capitalism we are dealing with an evolutionary process. It is by nature a form or method of economic change and not only never is but never can be stationary. The fundamental impulse that sets and keeps the capitalist engine in motion comes from new goods and new methods of production and transportation, created by the Schumpeterian entrepreneur, who is always on the outlook for new combinations of the factors of production (Heertje 2006).

The process of creative destruction takes time. For that reason there is no point in appraising its performance within a static framework. A system may produce an optimal allocation of resources at every point of time and may yet in the long run be inferior to a system without such optimal allocation, because the non-optimality may be a condition for the level and speed of long-run performances, in other words for dynamic efficiency. Furthermore, the process of creative destruction in Schumpeter's vision must be seen as the background for individual decisions and strategies. Economic theory has a tendency to concentrate on decisions about prices by firms, which are assumed to maximize profits, within a given structure. Schumpeter argues that the relevant problem is how capitalism creates and destroys these structures (Metcalfe 1998).

Schumpeter's conception of creative destruction overturns the idea that price competition is the only component of the market behaviour of entrepreneurs. In fact, it is not that kind of competition which counts, but the competition from the new commodity, the new technology, the new source of supply, and the new type of organization. Instead of marginal changes, fundamental upheavals are brought about by process and product innovations of existing firms and potential competitors.

Restrictive practices of monopolists and large firms are to be judged against the background of the perennial gale of creative destruction, rather than in the context of stationary development. The potential threat of process and product innovation reduces the scope and importance of restrictive practices that aim to guarantee the monopolist or big firm a quiet life. If however the profits are used to counterattack, restrictive practices may help to deepen the process of creative destruction and, therefore, the dynamic effects of capitalism (Reisman 2004).

The process of creative destruction as described by Schumpeter has been experienced again since the 1980s in the United States,

Japan, and Western Europe and since the 1990s in China and India as well. On the basis of new technologies many old firms, structures, and professions have been swept away, and new industrial organizations and labour relations have emerged. In particular, the application of information technology and the Internet with the dramatic decrease in transaction costs of communication is leading to major changes of a quantitative and qualitative nature in both the private and public sector of the economy. On the one hand, 'external' growth of already large firms which take over others is a feature of modern capitalism; on the other hand, every day new small firms are established, often created by former executives of existing (and long-lived) companies.

This extensive discussion of the process of creative destruction illustrates Schumpeter's strong emphasis on the supply side of the economy. It would be an interesting question to study the impact of the process of creative destruction on employment. My guess would be that, on balance, the process of creative destruction is more creative than destructive, not only with regard to employment but also concerning broader perspectives of growth and welfare. This may be one of the reasons why Schumpeter's work has had a lasting and ever-increasing influence on economic theory.

See Also

- ▶ Creative Destruction
- ▶ Market Structure
- ▶ Schumpeter, Joseph Alois (1883–1950)

Bibliography

Heertje, A. 2006. *Schumpeter on the economics of innovation and the development of capitalism*. Cheltenham: Edward Elgar.

Metcalfe, J.S. 1998. *Evolutionary economics and creative destruction*. London: Routledge.

Reisman, D. 2004. *Schumpeter's market*. Cheltenham: Edward Elgar.

Schumpeter, J. 1942. *Capitalism, socialism, and democracy*. New York: Harper.

Credit

Ernst Baltensperger

While the volume and complexity of credit transactions has grown immensely over the centuries, the act of credit extension and debt creation, or lending and borrowing, as such, is probably as old as human society. To extend credit means to transfer the property rights on a given object (e.g. a sum of money) in exchange for a claim on specified objects (e.g. certain sums of money) at specified points of time in the future. To take credit, or go into debt, is the other side of the coin. Credit and debt have always posed some special problems of understanding for economists, beyond those associated with the production, trade and consumption of 'ordinary' goods like wheat or cloth, or factors of production like labour services. There exists, of course, a wide array of different forms of credit contracts in today's economies. Classifications are customary; for example, according to types of debtors or creditors (domestic or foreign, public or private, etc.), length of contract duration, type of security put forward by the debtor, or the use of the loan by the borrower. However, this essay will attempt to concentrate on the essential features common to all or most groups of credit transactions, rather than enumerate and describe the differences between specific types and forms of credit.

The Economic Function of Credit

The credit market is essentially a market for intertemporal exchange. Something is given up in the present in exchange for something else in the future – or vice versa, if seen from the point of view of the borrower. The future 'repayment' typically includes a compensation in excess of the original 'payment'; that is, interest. The rate of interest represents the relative price in the market for intertemporal exchange.

The possibility of intertemporal exchanges allows market participants the realization of utility gains, just as voluntary exchange in general is mutually advantageous. The basic reason for this is that individuals are not normally indifferent about the distribution of their consumption over time but care about it. This notion of 'time preference' – used here in its most general and neutral sense, which does not necessarily imply a preference for present over future consumption – was first clearly formulated by Fisher (1930), who viewed *dated* consumption possibilities as the consumer's objects of choice; that is, as separate arguments of his utility function. This allowed the application of the standard tools of microeconomic analysis to problems of inter-temporal choice and proved to be the clue to a clear understanding and analytical treatment of credit and debt. Fisher's treatment still captures the essence of credit and the function it performs in the economy. The given time profile of income (endowments) faced by individuals will often not represent their most desired distribution of the given total consumption over time. The existence of a credit market (the possibility of intertemporal exchange) allows them to transfer a given stream into a preferred stream – either by anticipating future consumption via borrowing ('deficit units') or by transferring consumption into the future via saving and lending ('surplus units'). Transactions of this kind can be mutually advantageous, due to differences in endowments and/or differences in preferences between individuals.

Given real investment opportunities (capital accumulation), the existence of a credit market in general also allows the choice of superior investment decisions, ultimately leading to a higher level of utility. Thus the presence of a credit market, like any other market, permits a more efficient allocation of inputs and outputs, especially with respect to time.

This Fisherian view of the credit market makes clear that it constitutes part of the 'real' economy. That is, it performs a 'real' function by helping to determine the 'real' equilibrium of the economy and the levels of satisfaction reached by its members. It also makes clear that credit can play an important role even in a pure exchange economy with no production and capital formation, given sufficient divergence in individual tastes and/or endowments. On the other hand, production and capital formation can, in principle, take place without credit. Resources can be set aside and invested directly by their owners (the savers). If the owners have no taste or ability for administering these investments, they can, in principle, hire labour (managers) to perform this job (wage, or equity, contracts instead of credit, or debt, contracts). That is, alternative contractual arrangements allowing capital formation and production are available. Of course, credit (debt) contracts, on the one hand, and work (equity) contracts, on the other hand, differ with respect to the way in which risks are shared between the parties involved and with respect to their incentive effects, and a credit market will in general, as already pointed out, be helpful in achieving an efficient allocation of resources and, ultimately, consumption.

Credit and Budget Constraints

A basic question arising with any credit transaction concerns the mechanisms which ensure that the debtor will meet his future payment obligations. As soon as he has obtained his credit, the borrower has, in principle, a strong incentive to 'run off'. This is linked to the question of the appropriate formulation of budget constraints in the presence of credit. What limits credit demand and present consumption (and the incentive to cheat)? Obviously, a credit market can come into existence and survive only if there exist disciplining mechanisms which serve to prevent, or at least severely restrict, dishonest behaviour. Penalties of one sort or another must be in force, be it through legal provisions (bankruptcy laws), social stigmatization or simply the exclusion from, or discrimination in, future credit market participation.

The appropriate formulation of intertemporal budget constraints, in view of a credit market, is comparatively unproblematic (1) as long as the future payment capacity of a potential debtor (his future income stream) is known with perfect certainty, and (2) if, due to social institutions guaranteeing complete enforceability, there is

complete confidence in his willingness to fulfil his future payment obligations, as long as he objectively can. Under these conditions, the relevant magnitude serving to constrain an individual's lifetime consumption obviously is the present value of his lifetime income stream.

Matters are more complicated if the future is not perfectly foreseeable and/or contract enforceability is less than perfect. Unless credit extension is limited to the most pessimistic estimate of the debtor's future income or willingness to repay, there is then a possibility of default. Normally, creditors are willing to accept a certain positive probability of default in exchange for compensation in the form of a higher contractual rate of interest (a risk premium). However, the willingness to extend credit is affected, of course, by the possibility of default and its dependence on the amount of credit extended. Given a finite repayment capacity (finite future income), an increasing level of indebtedness increases the probability of default in two ways. First, for 'external' reasons: the possibility that the future payment obligations exceed the (uncertain) future repayment ability increases with increasing debt. Second, for 'internal' reasons (moral hazard): the incentive to 'run off' after credit has been obtained increases with an increasing repayment obligation; similarly, the incentive to produce future income may be lowered, since in case of partial default the debtor does not benefit from his own efforts. Given a finite repayment capacity, in fact, a point will be reached, sooner or later, where no increase in the contractual interest rate (no risk premium) can compensate the lender for the extra risk of non-payment resulting from a further increase in the level of debt, thus creating an absolute limit to the supply of credit to individuals. This was pointed out by Hodgman (1960), and has led him to speak of credit rationing.

An adequate level of trust in the implicit and explicit promises associated with outstanding debt contracts is an important prerequisite of a smoothly and efficiently operating financial system. Due to the intangible nature of 'trust', the danger of financial crises occurring whenever it is somehow weakened has always been inherent in a credit system. Institutional arrangements, such as a lender of last resort (usually the central bank) or an insurance system of one sort or another (e.g. deposit insurance) are important elements affecting the probability of such occurrences. They are traditionally seen as devices serving to eliminate, or at least contain, the risk of adverse chain reactions. Of course, one danger of institutions of this sort is that they may easily create a moral hazard problem themselves, by lowering the private costs of illiquidity and payment difficulties and thus reducing the private incentives to avoid excessive risks.

Imperfect Information and the Credit Market

In recent years the fact has been stressed that asymmetric information between market participants, and the resultant problems of adverse incentives and adverse selection, can lead to the breakdown of certain markets (incomplete markets) and to unusual types of market equilibria. These include equilibria with non-price rationing; that is, situations where the interest rate on a loan category is set by the lender at a given level and maintained there, even if there exists an excess demand for loans at this rate (Stiglitz and Weiss 1981). Starting from the notion that the lender, due to asymmetric information, must, to a certain degree, lump heterogeneous loan customers together, the basic idea is that an increase in the loan rate (applying equally to all customers) will induce 'good' (high quality) customers to leave and 'bad' (low quality) customers to stay (adverse selection), or that individual customers will be induced by the higher loan rate to choose riskier investment projects (moral hazard). In either case, the average quality of loan customers is reduced. Thus an increase in the loan rate here has, in addition to its usual positive effect on lender return, a negative effect which may possibly dominate the former. If this is the case, it is not in the interest of the lender to raise the loan rate, even in the face of an excess demand for loans. The loan rate has then lost its traditional allocative role of bringing in line supply and demand, and instead serves as a device to limit the damages resulting

from adverse selection and adverse incentives. Funds then must be allocated to customers in some other way.

This problem disappears again if creditors are able to overcome the underlying information asymmetries and identify different quality customers. Then they can offer different types of contracts (combinations of credit volumes and interest rates, possibly also of collateral levels and equity requirements) to different types of customers. One possibility which has been discussed, in analogy to similar problems in insurance and labour markets, concerns the feasibility of self-selection mechanisms. Under certain conditions it may be possible, by exploring the differences in preferences between high and low quality customers, to offer different types of contracts, so that each potential debtor has an incentive to choose of his own will the appropriate offer designed for his quality class. Another possibility concerns the ability of lenders to overcome the information deficiencies underlying the problems of adverse selection and incentives directly through information acquisition technologies of various sorts (direct screening and policing). Since this kind of information is customer-specific, this can encourage the development of long-term customer relationships. The empirical importance of the information-asymmetry models of credit-market behaviour referred to above thus will ultimately have to be judged in view of the empirical weight of these alternative response possibilities.

Credit and Credit Institutions

The role of credit as such must be clearly separated from the economic role of credit institutions, such as banks, playing the role of specialized intermediaries in the credit market by buying and simultaneously selling credit instruments (of a different type and quality). Since the ultimate borrowers and lenders can, in principle, do business with each other directly, without the help of such an intermediary, the function of these middlemen must be viewed as separate from that of credit as such.

Two main functions of institutions of this kind can be distinguished. The first is the function of risk consolidation or transformation. By dealing with a large number of creditors and debtors acting, to a considerable extent, independently of each other, the bank can, by exploiting the law of large numbers, achieve a consolidation of risks. In a world of subjective risk aversion, or if risk implies 'objective' costs of one sort or another (costs of adjusting to certain unfavourable states of the world), such a risk consolidation represents a utility gain for the individuals concerned, and this is a marketable service offered by these institutions to the public. Thus existence of risk and uncertainty (imperfect information) is fundamental for this first function of credit institutions.

The second major function of these institutions is that of a broker in the credit markets. As such, they specialize in producing intertemporal exchange transactions and owe their existence to their ability to bring together creditors and debtors at lower costs than the latter can achieve in direct transactions themselves. Transactions and information costs ('market imperfections') in the credit market, including the cost of evaluating credit risks as an especially important example, are fundamental for the financial intermediary in this second function. To summarize: the existence and function of credit institutions is linked in an essential way to the presence of uncertainty, imperfect information, and transactions costs in the credit market. In the absence of these elements, financial intermediaries would have no raison d'être (while credit as such can still perform an important function). Government, when issuing government bonds, can be viewed as an intermediary in a similar sense.

Another, basically similar, 'institutional' question concerns the marketability, or negotiability, of credit contracts and the existence of 'secondary' markets where they can be traded on a regular basis. This requires certain characteristics. In particular, the market cannot be too small, it must be comparatively homogeneous, and it must be possible to assess the quality of the traded contracts at reasonably low costs. The advantage to the

creditor of such a resale market is, of course, its contribution to the liquidity of these assets.

Credit in Macroeconomic Theory

In macroeconomic theory, the credit market has frequently played the role of the 'hidden' market eliminated from explicit consideration via application of Walras' Law. Although not explicitly appearing, a credit market (in the form of a bond market) is, however, present in most traditional macromodels. This was clearly brought out, in particular, by Patinkin (1956). Credit has traditionally played a prominent role in some specific issues of macroanalysis, nevertheless. In particular, this is the case with respect to the question of wealth effects. To what extent does credit creation represent creation of net wealth (and in turn affect aggregate demand)? This became one of the dominant issues in monetary theory and macroeconomics during the 1950s and 1960s. See, in particular, Patinkin (1956). Aggregate demand for goods (as well as for money and other assets) was seen as depending on aggregate net wealth of the private sector, in addition to income and relative prices, and all assets were examined with regard to the existence of an equivalent and offsetting liability within the private sector. For most financial assets, such an offsetting liability obviously exists. The exceptions, in the traditional view, were money and – with less confidence, because of the question of the capitalization of future tax liabilities required to finance interest payments – government bonds. As Niehans (1978, p. 91) has argued, this emphasis on net wealth was misplaced in the sense that it failed to appreciate that demand effects arising from individual components of wealth can be powerful even if net wealth effects are negligible or nonexistent. That is, it is not just net wealth which affects the demand for goods and assets; rather, the stocks of the various wealth components given at any point in time, and their difference from the corresponding long-run desired levels, determine the economy's attempts to build up or reduce these components over time.

Another macroeconomic area where the credit market has traditionally played an important role is money supply theory or, more generally, aggregate models of the financial sector of the economy (e.g. Brunner and Meltzer 1968; Tobin 1969). Credit markets and credit creation are seen in these models in the light of their relation to money markets and money creation and nominal (price level) control of the system. Financial markets here are typically disaggregated into markets for assets serving as media of exchange (government money and bank demand deposits) and other (non-money) assets, such as bonds and other similar credit instruments. Models of this type have helped considerably to clarify the role of central bank policies in controlling monetary aggregates and, ultimately, the price level. In particular, they have shown that, as long as the degree of substitutability between money and other assets is less than perfect, central bank control over a comparatively narrow monetary aggregate, such as base money, is sufficient for nominal control of the system (price level control), a large menu and volume of private credit notwithstanding.

See Also

► Financial Intermediaries

Bibliography

Brunner, K., and A.H. Meltzer. 1968. Liquidity traps for money, bank credit, and interest rates. *Journal of Political Economy* 76: 1–37.

Fisher, I. 1930. *The theory of interest*. New York: Macmillan.

Hodgman, D.R. 1960. Credit risk and credit rationing. *Quarterly Journal of Economics* 74(2): 258–278.

Niehans, J. 1978. Metzler, wealth, and macroeconomics: A review. *Journal of Economic Literature* 16(1): 84–95.

Patinkin, D. 1956. *Money, interest, and prices*. Evanston: Row & Peterson, 2nd ed. New York: Harper & Row, 1965.

Stiglitz, J., and A. Weiss. 1981. Credit rationing in markets with imperfect information. *American Economic Review* 71(3): 393–410.

Tobin, J. 1969. A general equilibrium approach to monetary theory. *Journal of Money, Credit, and Banking* 1(1): 15–29.

Credit Card Industry

Victor Stango and Julian Wright

Keywords

Credit card industry; Interchange fees; Interest rates; Stickiness of; Networks; Sticky prices; Two-sided markets

JEL Classifications

L89

The concept of a general purpose credit card originated in 1949, when Frank McNamara dined in a New York restaurant and discovered that he could not pay for his meal (Evans and Schmalensee 1999). By the 1980s credit cards had become ubiquitous, and they remain a popular form of payment in most economies. Banks offer cards, setting terms such as interest rates and annual fees. Transactions are handled by networks such as Visa and MasterCard, which emerged in the 1970s as joint member associations. Early research examining the market typically focused on the retail level, while more recent work has tended to focus on the network level, mirroring a shift in policy concerns in the 1980s.

In its early years the US retail credit card market was characterized by extreme interest rate 'stickiness' – credit card rates remained virtually constant over time, regardless of economy-wide changes in interest rates. Credit card issuers also appear to have earned super-normal profits during the same period. This presents a puzzle in an industry displaying many classic characteristics of a perfectly competitive market (Ausubel 1991). Ausubel suggests a variety of explanations for this puzzle, including the possibility that credit card borrowers do not fully anticipate the degree to which they will use the cards.

Ausubel's research spurred a wave of subsequent work proposing explanations for interest rate stickiness. Mester (1994) and Brito and Hartley (1995) provide theoretical explanations for interest rate stickiness based on asymmetric information or consumer transaction costs. Calem and Mester (1995) provide empirical evidence that consumer search and switching costs might explain interest rate stickiness. A complementary explanation for interest rate stickiness is that state-level interest rate ceilings during the 1980s facilitated tacit collusion among card issuers, leading to greater-than-normal interest rate stability (Knittel and Stango 2003).

By the early 1990s interest rates had become much more flexible as credit card issuers switched to variable interest rates. By most accounts, the market also became more competitive during this time. One explanation for the change is technological progress that allowed more efficient credit scoring by large nationally marketed card issuers, creating a truly national market that fostered aggressive competition. Other explanations include the threat of interest rate regulation and the entry of new issuers.

At the network level, the key economic issue is that payment card systems like MasterCard and Visa are two-sided markets: they have to attract cardholders to get merchants and merchants to get cardholders. Diners Club did this in 1950 by initially giving away cards to consumers and charging merchants seven per cent of their bill. These days, consumers obtain rewards for using their cards. This structure of pricing has raised the concern of some policymakers. In their view, retailers pay too much to accept credit cards, costs that end up being covered by consumers who do not use credit cards (by way of higher retail prices). Card associations sustain such a price structure through the setting of an interchange fee, which determines how much the merchant's bank must pay the cardholder's bank for each card transaction. A high interchange fee results in a high merchant fee and a low (or negative) fee for cardholders.

The issue of how much to charge each type of user is a common one in other two-sided markets. Magazines and newspapers decide how much to charge readers versus advertisers, and shopping malls decide how much to charge shoppers versus shops. The interest of policymakers in credit cards has spurred research in two-sided markets more generally.

Baxter (1983) provides an early analysis of interchange fees (see Rochet 2003, for a survey). His key insight is that efficiency calls for card transactions whenever the *joint* benefits to the consumer and merchant of using the card exceed the joint costs of doing so. In the absence of an interchange fee, each type of user will face only the private costs and benefits of cards. A payment from the merchant's bank (acquirer) to the cardholder's bank (issuer) via the interchange fee can align the private incentive to use cards with the social incentive. This provides a justification for setting an interchange fee, but does not imply that card associations will set it at the right level.

One reason a card association might set the interchange fee too high is that acquirers may pass through a larger proportion of interchange fees into merchant fees than issuers pass back to cardholders (in the form of lower fees or higher rewards). Then associations will want to pass revenues to the issuing side, via high interchange fees, where they are competed away less aggressively. A second possible reason is that, if merchants accept cards to attract customers from each other, their private willingness to accept cards includes the surplus their customers get from using cards. As a result, cardholder surplus is over-represented, and card associations tend to charge merchants too much and cardholders too little. Although these theoretical possibilities highlight possible divergences between privately and socially optimal interchange fees, they provide no basis for the cost-based regulation of interchange fees.

See Also

► Two-Sided Markets

Bibliography

Ausubel, L. 1991. The failure of competition in the credit card market. *American Economic Review* 81: 50–81.

Baxter, W. 1983. Bank interchange of transactional paper: Legal perspectives. *Journal of Law and Economics* 26: 541–588.

Brito, D., and P. Hartley. 1995. Consumer rationality and credit cards. *Journal of Political Economy* 103: 400–433.

Calem, P., and L. Mester. 1995. Consumer behavior and the stickiness of credit-card interest rates. *American Economic Review* 85: 1327–1336.

Evans, D., and R. Schmalensee. 1999. *Paying with plastic*. Cambridge, MA: MIT Press.

Knittel, C., and V. Stango. 2003. Price ceilings as focal points for tacit collusion: Evidence from credit cards. *American Economic Review* 93: 1703–1729.

Mester, L. 1994. Why are credit card rates sticky? *Economic Theory* 4: 505–530.

Rochet, J.-C. 2003. The theory of interchange fees: A synthesis of recent contributions. *Review of Network Economics* 2(2): 97–124.

Credit Crunch Chronology: April 2007–September 2009

Barry Turner

Abstract

The global financial crisis that began in mid-2007 and exploded in the fall of 2008 shocked most economists. Some had raised concerns about the rapid growth in the housing market in developed countries, especially to "sub-prime," high-risk borrowers. Others had been concerned about large banks being "Too Big to Fail," worrying that such banks might take inordinate risk since they had an implicit government backstop. But the typical economist–even the typical macroeconomic forecaster–was not predicting a massive global recession over the 2007–2008 period. Thus, the crisis was a genuine surprise.

While economists have theories to help explain and understand recessions, bubbles, manias and crashes, only by taking these theories to the data will we learn which models are relevant and which are mere theoretical curiosities. The chronology below should help refresh reader's memories about the world-shaking events surrounding the crisis while also reminding them of some of less-famous but possibly still crucial moments from the 2007 to 2009 period. Many of the events,

institutions, and concepts below are discussed in full-length articles elsewhere in the Dictionary.

Keywords

Global financial crisis; History; Great recession; Subprime mortgage crisis; Banking crisis

JEL Classifications

N12; N14; N22; N24; E44

April 2007

2nd – New Century Financial, based in California and second only to HSBC in the US sub-prime mortgage market, filed for Chapter 11 bankruptcy protection, making over 3,200 employees redundant.

May 2007

3rd – Dillon Read Capital Management, a hedge fund, was forced to shut down following a SFr150m. (US$123 m.) first-quarter loss on US sub-prime mortgage investments.

June 2007

25th – Queen's Walk Investment announced a loss of €67.7 m. (US$91 m.) in the year ending 31 March, reflecting a decline in the value of its UK and US mortgage-linked securities holdings.

28th – Caliber Global Investment, a London-listed fund, announced it would wind down over twelve months following a £4.4 m. (US$8.8 m.) loss from sub-prime investments.

29th – US investment bank Bear Stearns replaced the chairman and chief executive of its asset management business in an effort to restore investor confidence following the collapse of two of its hedge funds invested in the sub-prime mortgage market.

July 2007

3rd – United Capital Asset Management, a Florida-based hedge fund, suspended investor redemptions following heavy losses in sub-prime bonds and derivatives.

11th – Braddock Financial, based in Denver, Colorado closed its US$300 m. Galena fund owing to sub-prime losses.

19th – Ben Bernanke, chairman of the Federal Reserve, warned that the sub-prime crisis in the USA could cost up to US$100bn.

27th – Absolute Capital, an Australian hedge fund, temporarily suspended redemptions for two of its funds.

31st – After losing over 50% of its capital, Boston-based hedge fund, Sowood Capital Management, was bought by larger rival, Citadel.

August 2007

1st – Shares in Australia's Macquarie Bank fell by more than 10% after a warning to investors that its two Fortress funds could lose more than $A300m. (US$256 m.).

1st – Bear Stearns halted redemptions in a third hedge fund, Asset-Backed Securities, following a rush of withdrawals.

1st – German bank IKB was bailed out by rival banks for h8bn. after it was exposed to losses in the US sub-prime sector.

6th – American Home Mortgage Investment (AHM), the tenth biggest home loan lender in the USA, filed for Chapter 11 bankruptcy protection.

9th – France's largest bank, BNP Paribas, suspended three of its funds exposed to the US sub-prime mortgage market.

9th – The European Central Bank (ECB) injected €94.8bn. into the eurozone banking market to stabilize overnight interest rates. The Fed quickly followed the ECB by announcing that it would provide US$12bn. of temporary reserves to the American banking system.

10th – Continuing turmoil in the markets forced action from the world's central banks. In total US$120bn. of extra liquidity was pumped into financial markets.

10th – The FTSE 100 Index fell by 3.7%, its largest drop in four years.

13th – Investment bank Goldman Sachs injected US$3bn. into its Global Equity Opportunities hedge fund.

16th – The USA's largest mortgage lender, Countrywide Financial, received an US$11.5bn. lifeline from 40 of the world's largest banks.

17th – The US Federal Reserve cut its primary discount rate, the rate at which it lends money to banks, by half a point from 6.25% to 5.75%.

22nd – Countrywide Financial received a US$2bn. capital injection from the Bank of America.

23rd – US and European banks, including the Bank of America, Citigroup, JP Morgan Chase and Germany's Deutsche Bank, borrowed US$2bn. from the US Federal Reserve to improve credit access.

23rd – Lehman Brothers closed its sub-prime mortgage unit, BNC Mortgage, releasing 1,200 workers.

31st – President George W. Bush announced plans to help struggling sub-prime mortgage borrowers. Federal Reserve chairman Ben Bernanke pledged to take action to protect the wider economy from market turmoil.

September 2007

6th – The US Federal Reserve added US$31.25bn. to the US money markets and the ECB lent an extra €42.2bn. to banks.

10th – Victoria Mortgages, owned by US private equity group Venturion Capital, was forced into administration, becoming the first UK casualty of the sub-prime crisis.

13th – The Bank of England provided emergency financial support to Northern Rock, the UK's fifth largest mortgage lender.

17th – UK Chancellor Alistair Darling guaranteed Northern Rock's savings accounts, following several days of a run on the bank's deposits.

18th – The US Federal Reserve cut interest rates by half a point from 5.25% to 4.75%.

20th – Goldman Sachs announced record profits after hedging that the value of mortgage bonds would fall, despite losing US$1.5bn. from the sub-prime crisis.

26th – UK banks shunned the Bank of England's auction of £10bn. worth of three-month loans, an emergency funding facility introduced by Governor Mervyn King.

October 2007

1st – Swiss bank UBS revealed a writedown of SFr4bn. (US$3.4bn.) on hedge fund losses and exposure to the sub-prime mortgage market. The group announced plans to shed 1,500 jobs.

5th – Investment bank Merrill Lynch revealed a third-quarter writedown of US$5.5bn.

15th – Citigroup announced a total of US$6.5bn. in writedowns.

24th – Merrill Lynch announced US$8.4bn. of losses and writedowns. A quarterly loss of US$2.24bn. was the largest in its history. Stan O'Neal, chief executive, resigned six days later.

31st – The US Federal Reserve reduced interest rates from 4.75% to 4.5%.

November 2007

1st – Swiss bank Credit Suisse revealed a US$1bn. writedown.

4th – Citigroup announced further writedowns of US$8–11bn. Charles Prince resigned as chairman and chief executive.

7th – US investment bank Morgan Stanley forecast a loss of US$3.7bn. against fourth-quarter revenues.

9th – Wachovia, the USA's fourth largest lender, unveiled losses of US$1.1bn. for Oct. owing to the continued decline in value of its mortgage debt.

13th – The Bank of America revealed it would write off US$3bn. of bad debts linked to the US sub-prime crisis during the last quarter of 2007 and would inject a further US$600 m. into a structured investment vehicle with high exposure to sub-prime mortgages.

14th – HSBC, the world's second largest bank, claimed it was writing off US$38 m. of loans a day to struggling Americans and raising its sub-prime bad debt provision to US$3.4bn.

14th – The Bank of England forecast a sharp slowdown in UK domestic growth in 2008 together with higher inflation.

15th – Barclays, the UK's third largest bank, announced a writedown of US$2.6bn. on securities related to the US sub-prime mortgage market, having lost US$1.64bn. in Oct. alone.

16th – Northern Rock's Adam Applegarth resigned as chief executive.

20th – Shares in Paragon, the UK's third largest buy-to-let mortgage lender, were suspended after falling in value by 50%. It warned shareholders it could face collapse if it could not raise an extra £250m.

20th – Freddie Mac, the USA's second largest provider of mortgage financing, announced its largest quarterly loss so far after unveiling US$4.8bn. of bad debts and writedowns.

27th – Citigroup agreed to sell shares in its company worth US$7.5bn. to the Abu Dhabi Investment Authority, making it the largest shareholder with a stake of 4.9%.

December 2007

4th – The Bank of Canada cut interest rates by a quarter of a percentage point from 4.5% to 4.25%.

6th – The Bank of England lowered interest rates, from 5.75% to 5.5%.

6th – RBS warned investors it expected to write off £1.25bn. as a result of exposure to the US sub-prime mortgage market.

6th – President Bush unveiled plans to freeze rates on sub-prime mortgages for the next five years.

10th – UBS revealed it had written off a further SFr11.2bn. (US$10bn.) against its US sub-prime mortgage exposure.

10th – France's second largest bank, Société Générale, moved to bailout its structured investment vehicle with a credit line of up to US$4.3bn.

11th – The US Federal Reserve cut interest rates for the third time in four months, reducing them from 4.5% to 4.25%.

12th – Five central banks from the UK, Europe and USA launched a US$110bn. joint cash injection targeting international interbank borrowing markets.

14th – Citigroup brought US$49bn. worth of sub-prime debts to keep afloat seven high-risk structured investment vehicles.

17th – The US Federal Reserve made US$20bn. available to major banks to ease interbank lending rates as the first part of a plan agreed by five central banks.

18th – The Bank of England released d10bn. of funds to UK banks and financial institutions.

18th – The ECB injected €348.7bn. (US$502bn.) into banks to help ease credit fears over the Christmas period.

19th – US investment bank Morgan Stanley wrote down US$9.4bn. in sub-prime losses. A cash injection of US$5bn. (equating to 9.9% of the bank) was provided by China Investment Corporation (CIC).

January 2008

9th – The World Bank forecast a 0.3% slowdown in global economic growth to 3.3% in 2008 but claimed growth in China and India would soften the impact.

9th – James Cayne, chief executive of US investment bank Bear Stearns, stepped down.

11th – Countrywide Financial, the USA's largest mortgage lender, was bought by the Bank of America for US$4bn.

15th – Citigroup reported a US$9.8bn. loss for the fourth quarter, the largest in its history. The bank also announced a capital injection of US$6.9bn. from the Government of Singapore Investment Corporation (GIC). In total Citigroup and Merrill Lynch had received over US$21bn. from foreign investors including Saudi Arabia and Kuwait.

21st – Stock markets across the world suffered their biggest losses since 11 Sept. 2001, triggered by fears of a looming recession in the USA.

22nd – The US Federal Reserve slashed interest rates by 0.75% to 3.5%, its largest cut in over 25 years.

28th – European bank Fortis warned that its losses connected to US sub-prime mortgage debt could be as much as €1bn. (US$1.5bn.).

30th – The US Federal Reserve cut interest rates by a further 50 basis points from 3.5% to 3.0%.

31st – MBIA, the world's largest bond insurer, revealed a US$2.3bn. loss in the fourth quarter.

February 2008

6th – Wall Street had its worst share losses in over a year, while the UK's FTSE 100 fell by 2.6%.

7th – The Bank of England reduced interest rates from 5.5% to 5.25%.

10th – Finance ministers from the G7 group of industrialized nations warned of worldwide losses from the US mortgage crisis of up to US$400bn.

13th – The Financial Services Agency, Japan's financial watchdog, said Japanese banks had lost a total of 600bn. yen (US$5.6bn.) from the US sub-prime mortgage crisis in the previous 12 months.

14th – UBS confirmed it had made a loss of SFr4.4bn. (US$4bn.) in 2007, following US$18.4bn. of writedowns.

14th – Commerzbank, Germany's second largest bank, announced writedowns of €774m. (US$1.1bn.), despite record-year profits.

17th – UK Chancellor Alistair Darling confirmed mortgage lender Northern Rock would be brought into temporary public ownership.

March 2008

3rd – HSBC, the UK's largest bank, unveiled total writedowns of US$17.2bn., despite an annual profit increase of 10%.

5th – Credit Agricole, France's largest retail bank, announced a loss of €857m. (US$1.3bn.) in the fourth quarter, following a €3.3bn. charge at its Calyon investment banking arm on losses related to the credit crisis.

6th – Peloton Partners, a London-based hedge fund, was forced to liquidate its £1bn. ABS Master Fund after failing to meet interest payments on loans taken out to buy assets.

7th – Carlyle Capital Corporation, a US$22bn. credit fund owned by US private equity firm Carlyle Group, collapsed.

7th – The former chief executives of Merrill Lynch, Citigroup and Countrywide Financial were questioned before a Congressional committee over their large salary and pay-off packages while their firms experienced heavy losses.

7th – The US Federal Reserve made available up to US$200bn. of emergency financing in response to 'rapid deterioration' in the credit markets.

14th – US investment bank Bear Stearns received emergency funding from JP Morgan Chase with the US Federal Reserve's backing, following a collapse in confidence from its hedge fund clients.

16th – Bear Stearns was bought out by JP Morgan Chase for US$236 m or US$2 per share, a fraction of its previous value, backed by US$30bn. in loans from the US Federal Reserve.

16th – The US Federal Reserve lowered its lending rate to financial institutions by a quarter of a point to 3.25% and created a new lending facility for large investment banks to secure short-term loans.

18th – Wall Street investment banks Goldman Sachs and Lehman Brothers reported a halving of profits in the first quarter of 2008. The results were better than expected, boosting shares in both firms.

31st – Henry Paulson, the US Treasury Secretary, announced a package of reforms designed to help the Federal Reserve tackle financial market turmoil and improve regulation of the financial system.

April 2008

1st – UBS revealed a further US$19bn. of asset writedowns on top of the US$18.4bn. already lost in 2007. Chief executive Marcel Ospel resigned.

7th – UK mortgage lender Abbey withdrew 100% mortgage deals available to UK borrowers.

8th – The IMF warned potential losses from the global credit crunch could reach US$945bn.

10th – The Bank of England cut interest rates by a quarter point to 5%.

14th – Wachovia, the fourth largest US bank, revealed a US$4.4bn. writedown for the first quarter following a jump in foreclosures in California and Florida.

16th – JP Morgan Chase reported a US$5.1bn. writedown for the first quarter against investments in mortgage-backed securities and its portfolio of homeloans.

17th – Merrill Lynch unveiled a loss of US$1.96bn. in the first quarter.

18th – Citigroup posted its second consecutive quarterly loss, of US$5.1bn., and announced it would cut 9,000 jobs after writing off US$15.1bn. in toxic assets.

21st – The Bank of England unveiled a £50bn. plan to aid the UK banks by allowing lenders to exchange potentially risky mortgage debts for government-backed bonds.

22nd – RBS, the UK's second largest bank, revealed pre-tax writedowns of £5.9bn. and requested £12bn. from shareholders to rebuild its capital base.

24th – Credit Suisse reported a quarterly loss of SFr2.5bn. (US$2.1bn.), its first loss in nearly five years, following asset writedowns of US$5.2bn.

30th – Nationwide Building Society recorded the first annual fall in UK house prices for ten years, with prices 1% lower in April than the previous year.

May 2008

2nd – The US Federal Reserve, European Central Bank and Swiss National Bank expanded liquidity by injecting an extra US$82bn. into the banking system.

12th – HSBC announced it had written off US$3.2bn. in the first quarter as a result of the sub-prime crisis.

13th – UK bank Alliance & Leicester disclosed a £391m. writedown in the first quarter.

14th – UK mortgage lender Bradford & Bingley launched an emergency £300m. rights issue.

15th – Barclays revealed a further £1.7bn. in writedowns.

22nd – Swiss bank UBS launched a SFr16bn. (US$15.5bn.) rights issue to cover its US$37bn. writedowns.

June 2008

19th – Chicago-based firm Hedge Fund Research showed 170 funds had been forced into liquidation during the first quarter, while fewer funds were launched than at any time since 2000.

19th – Two former managers of US investment bank Bear Stearns were charged with fraud. It was alleged they had misled investors about the health of their hedge funds.

25th – Major new investors in Barclays, including the Qatar Investment Authority, invested £1.7bn. (US$3.3bn.) for a 7.7% share in the business.

July 2008

8th – A quarterly survey of businesses by the British Chambers of Commerce (BCC) found that the UK faced a serious risk of recession.

10th – Share prices in the USA's two largest mortgage finance companies, Fannie Mae and Freddie Mac, plummeted by nearly 50% as investor anxiety grew over government intervention that would leave their stock worthless.

11th – The FTSE 100 fell deep into a bear market (a 20% fall from its market peak in June 2007) as blue-chip stocks reached their lowest level since 31 Oct. 2005.

13th – US mortgage lender IndyMac Bank, based in California, collapsed, becoming the second largest financial institution to fall in US history.

14th – The US government announced emergency measures to expand credit access to mortgage finance companies Fannie Mae and Freddie Mac, and allow the Treasury to buy shares in the companies.

30th – UK bank Lloyds TSB revealed £585m. of writedowns as pre-tax profits fell by 70% in the first half of the year.

31st – Nationwide recorded an 8.1% fall in the value of houses, the biggest annual fall in UK house prices since their surveys began in 1991.

31st – Halifax Bank of Scotland (HBOS) announced that its first-half profits fell by 72% to £848m. while bad debts rose by 36% to £1.31bn.

August 2008

1st – UK mortgage lender Alliance & Leicester revealed a £209m. hit on risky assets and higher funding costs as pre-tax profits for the first half of the year fell by 99% on the previous year.

1st – US mortgage lender IndyMac Bank filed for Chapter 7 bankruptcy protection.

4th – HSBC announced a 28% decline in half-year profits to £5.1bn.

5th – French bank Société Générale reported a 63% fall in second-quarter profits, after its investment banking division lost €1.2bn. (US$1.9bn.) from sub-prime related investments.

6th – US mortgage lender Freddie Mac announced a second quarter loss of US$822 m., its fourth successive loss, with credit-related expenses doubling to US$2.8bn. and US$1bn. lost on company writedowns on the value of sub-prime mortgages.

7th – Barclays revealed a 33% decline in first-half year profits together with further writedowns of £2.4bn. from bad loans and other credit impairment charges.

8th – RBS announced the second largest loss in UK banking history, with a pre-tax loss of £692m. for the first half of the year, resulting from £5.9bn. of writedowns.

29th – UK mortgage lender Bradford & Bingley reported a loss of £26.7 m for the first six months of the year.

30th – Chancellor Alistair Darling warned that the UK economy faced its worst economic crisis in 60 years and claimed that the downturn would be more 'profound and long-lasting' than most people had imagined.

September 2008

5th – Fears over a global economic slowdown, combined with news that the US economy had shed 84,000 jobs the previous month, led to losses in global stock markets. London's FTSE 100 experienced its biggest weekly decline since July 2002, while markets in Paris, Frankfurt, Japan, Hong Kong, China, Australia and India all fell between 2 and 3%.

7th – US mortgage lenders Fannie Mae and Freddie Mac, who together accounted for nearly half of all outstanding mortgages in the USA, were taken into public ownership in one of the largest bail-outs in US history.

7th – In the UK, Nationwide Building Society took ownership of smaller rivals Derbyshire and Cheshire Building Societies.

10th – The European Commission predicted that the UK, Spain and Germany would fall into recession and eurozone growth would fall to 1.3% in 2008, 0.4% less than previous projections.

15th – US investment bank Lehman Brothers filed for Chapter 11 bankruptcy protection after it was unable to find a buyer. It became the first major bank to collapse since the beginning of the credit crisis.

15th – The Bank of America bought out US bank Merrill Lynch for US$50bn.

15th – Fears over the strength of the global financial system following the collapse of Lehman Brothers caused stock markets across the globe to tumble. The FTSE 100 Index fell by 212.5 points, wiping d50bn. off the top 100 British companies, while the Dow Jones Industrial Average shed 504 points, its biggest fall since the 9/11 attacks.

16th – The US Federal Reserve launched an US$85bn. rescue package for AIG, America's largest insurance company, to protect it from bankruptcy in return for an 80% public stake in the business.

17th – Lloyds TSB agreed to take over HBOS, Britain's largest mortgage lender, in a deal worth d12bn. following a run on HBOS shares.

17th – UK bank Barclays bought Lehman Brothers' North American investment banking and trading unit for US$250 m., along with the company's New York HQ and two data centers for a further US$1.5bn.

18th – The US Federal Reserve, together with the European Central Bank, the Bank of England, the Bank of Japan, the Bank of Canada and the Swiss National Bank, pumped US$180bn. of extra liquidity into global money markets.

22nd – Japan's largest brokerage house Nomura Holdings Ltd acquired the Asian operations of Lehman Brothers, worth around US$230 m.

22nd – Wall Street banks Morgan Stanley and Goldman Sachs give up their status as investment banks to become lower risk, tightly regulated commercial banks.

23rd – Nomura Holdings acquired the European and Middle Eastern equities and investment banking operations of Lehman Brothers.

25th – US mortgage lender Washington Mutual collapsed. Its assets were sold to JP Morgan Chase for US$1.9bn.

25th – Ireland became the first eurozone economy to fall into recession.

29th – European bank Fortis was partially nationalized following talks between the European Central Bank and the Netherlands, Belgium and Luxembourg. Each country agreed to put €11.2bn. (US$16.1bn.) into the bank.

29th – UK mortgage lender Bradford & Bingley was taken into public ownership, with the government taking control of the company's £50bn. mortgages and loans, while its savings unit and branches were to be sold to Spain's Santander.

29th – US bank Wachovia agreed to a rescue takeover by Citigroup, absorbing US$42bn. of the company's losses.

29th – The Icelandic government took a 75% stake in Glitner, Iceland's third largest bank, for €600m. (US$860 m.).

29th – The German government injected €35bn. (US$50.2bn.) into Hypo Real Estate, the country's second largest commercial property lender.

29th – A US$700bn. rescue package was rejected by the US House of Representatives. Wall Street stocks plummeted, with the Dow Jones Index shedding 778 points, its biggest ever one-day fall. The FTSE 100 lost 269 points in one of its worst-ever trading days.

30th – European bank Dexia was bailed out, with the Belgian, French and Luxembourg governments injecting €6.4bn. (US$9bn.).

30th – The Irish government stepped in with €400bn. (US$562.5bn.) to guarantee all deposits, debts and bonds in six banks until September 2010.

30th – Japan's Nikkei 225 stock fell by 4.1% to register its lowest closing point since June 2005, while in Hong Kong the Hang Seng index ended the day down 2.4%.

October 2008

3rd – The US House of Representatives passed a US$700bn. rescue package. The plan aimed to buy up bad debts of failing banks while guaranteeing deposit accounts up to US$250,000.

3rd – US bank Wells Fargo announced a buy-out of Wachovia for US$15.1bn. *3rd* – The UK government increased guarantees for bank deposits to £50,000, effective from 7 October 2008.

6th – Germany's finance ministry, together with private banks, agreed a €50bn. (US$68bn.) deal to save Hypo Real Estate.

6th – French bank BNP Paribas announced it had agreed to take control of Fortis' operations in Belgium and Luxembourg, together with its international banking franchises, for €14.5bn. (US$19.7bn.).

6th – The Iceland Stock Exchange temporarily suspended trading in six of the economy's largest financial firms. Banks agreed to sell off their foreign assets to help bolster the domestic banking sector.

7th – The Icelandic government took control of Landsbanki, the nation's second largest bank. Internet bank Icesave, owned by Landsbanki, suspended all deposits and withdrawals.

8th – The UK government announced a £400bn. (US$692bn.) package of reforms, including £50bn. to the top eight financial institutions, an extra d100bn. available in short-term loans from the Bank of England and d250bn. in loan guarantees to encourage banks to lend to each other.

8th – Six central banks – the US Federal Reserve, the Bank of England, the European Central Bank, the Bank of Canada, the Swiss National Bank and Sveriges Riksbank – coordinated an

emergency interest rate cut of half a percentage point.

8th – The UK government announced that it planned to sue Iceland to recover deposits in Icesave, the failed Internet bank that had earlier stopped customers from withdrawing money.

9th – The IMF drew up emergency plans to make funds available to governments affected by the financial crisis.

10th – Japan's Nikkei stock average shed 881 points, or 9.62%, to fall to its lowest level since May 2003. Yamato Life Insurance became Japan's first major victim of the global financial crisis.

10th – Singapore officially fell into recession after the export-dependent economy experienced a fall in demand from US and European markets.

10th – The FTSE 100 closed down 8.85%, having lost 381.7 points, its worst fall since the crash of 1987, knocking £89.5bn. off the value of the UK's largest companies.

11th – The G7 nations agreed a five-point plan to unfreeze credit markets, including adoption of Britain's proposal to part-nationalize banks.

13th – The UK government announced an injection of £37bn. into RBS, Lloyds TSB and HBOS in return for a controlling share of each company.

13th – Germany and France led a coordinated plan to restore liquidity into their banking sectors in a move costing up to h2trn. for the EU's 27 states.

13th – The Dow Jones Industrial Average gained 936 points or 11%, its highest one-day gain and its largest percentage jump since 1933, following news of plans to increase bank liquidity.

14th – The US government revealed a US$250bn. plan to part-nationalize several banks.

15th – Retail sales in the US in Sept. recorded their biggest decline in over three years as the Dow Jones index fell by 7.87%, its largest decline since 26 Oct. 1987.

15th – JP Morgan Chase announced a quarterly profit fall of 84%, while Wells Fargo suffered a 25% drop in earnings.

16th – The Swiss government injected US$60bn. into UBS in return for a 9.3% stake and a boost in capital, while Credit Suisse turned down the offer of state aid but raised capital from private investors and a sovereign wealth fund.

16th – Citigroup posted its fourth consecutive quarterly loss with a shortfall of US$2.81bn. for the third quarter, following over US$13bn. of writedowns.

17th – French bank Caisse d'Epargne admitted a €600m. (US$807 m.) derivatives trading loss triggered by 'extreme market volatility' during the week of 6 October.

19th – Dutch savings bank ING received a €10bn. (US$13.4bn.) capital injection from the Netherlands authorities in return for preference shares in the company. The Dutch government established a €20bn. fund to support domestic banks as required.

19th – South Korea announced a rescue package worth US$130bn. offering a state guarantee on banks' foreign debts and promising liquidity to firms.

20th – Sweden's government offered credit guarantees up to 1.5trn. kroner (US$205bn.), with 15bn. kroner set aside in a bank stabilization fund.

22nd – US bank Wachovia reported a US$24bn. loss for the third quarter, the biggest quarterly loss of any bank since the beginning of the credit crunch.

24th – Official data showed that the UK economy contracted for the first time in 16 years, with a fall in economic growth of 0.5% for the third quarter.

24th – The Danish central bank raised interest rates by a half-point to 5.5%.

29th – The US Federal Reserve slashed interest rates by a half-point to 1%, its lowest level since June 2004.

29th – The IMF, European Union and World Bank announced a rescue package for Hungary, pledging US$25.1bn. to promote confidence in the country's financial markets and its currency.

30th – Deutsche Bank reported a large fall in profits following writedowns of €1.3bn. in the third quarter.

30th – Japan unveiled a 27trn. yen (US$270.6bn.) stimulus package for small businesses and to provide emergency cash to families exposed to the credit crunch.

31st – The Bank of Japan cut interest rates, from 0.5% to 0.3%, for the first time in seven years in response to the global financial crisis.

November 2008

4th – HBOS revealed writedowns for the nine months up to Sept. at £5.2bn., up from £2.7bn. for the first half of the year.

5th – The Italian government offered up to €30bn. (US$39bn.) to recapitalize banks.

5th – Australia's central bank slashed interest rates by a higher-than-expected 75 basis points to 5.25%, the lowest level since March 2005.

6th – The IMF approved a US$16.4bn. loan to Ukraine.

6th – The Bank of England reduced interest rates by 1.5% to 3%, the lowest level since 1955.

6th – The European Central Bank lowered interest rates by a half-point to 3.25%.

9th – The Chinese government announced a US$586bn. stimulus package. The plan to relax credit conditions, cut taxes and invest in infrastructure and social projects over a two-year period equated to 7% of the country's GDP.

11th – US electronics retailer Circuit City filed for Chapter 11 bankruptcy protection. It became the largest US retailer to fall victim to the credit crisis.

11th – Swedish investment bank Carnegie was taken over by the Swedish government after its license was revoked for failures in internal controls.

14th – The eurozone officially slipped into recession after figures showed the area shrunk by 0.2% for the second consecutive quarter.

20th – The IMF approved a US$2.1bn. loan for Iceland in an attempt to 'restore confidence and stabilize the economy.'

23rd – The US government agreed a bailout of Citigroup, injecting US$20bn. of capital in return for preference shares. The move included a guarantee of up to US$306bn. of Citigroup's risky loans and securities.

24th – In his pre-Budget report, Chancellor Alistair Darling unveiled a fiscal stimulus plan. VAT was reduced to 15% from 17.5% and an extra £20bn. was to be pumped into the economy, with government borrowing set to increase to record levels.

25th – The IMF approved a US$7.6bn. loan to Pakistan.

25th – The US Federal Reserve pumped a further US$800bn. into the economy, with US$600bn. to buy up mortgage-backed securities and US$200bn. to unfreeze the consumer credit market.

26th – The European Commission unveiled a €200bn. (US$256bn.) economic recovery plan.

December 2008

4th – French President Nicolas Sarkozy announced a €26bn. (US$33bn.) stimulus plan, including a €1bn. loan to carmakers and €5bn. of new public sector investments. The French government would offer companies €11.5bn. worth of credits and tax breaks on investments for 2009.

4th – The Bank of England cut interest rates by 1% to 2% with business surveys suggesting that the downturn had gathered pace.

4th – The Reserve Bank of New Zealand reduced interest rates by a record 150 basis points to 5%.

4th – The European Central Bank reduced its main interest rate by 75 basis points to 2.5%, its largest ever cut.

4th – Sweden's central bank cut interest rates by a record 1.75% to 2%, while Denmark's central bank Nationalbank followed with a 75 basis point reduction to 4.25%.

9th – The Bank of Canada lowered its benchmark interest rate by 75 basis points to 1.5%, its lowest rate since 1958.

11th – The Bank of Korea reduced interest rates by a record 1% to 3%.

16th – The US Federal Reserve slashed interest rates from 1% to a range between zero and 0.25%, its lowest recorded level.

19th – Japan's central bank cut interest rates from 0.3% to 0.1%, having projected that the economy would shrink by 0.8% in the current fiscal year and experience zero growth for the year ending March 2010.

19th – The US government pledged US$17.4bn. of its US$700bn. originally allocated for the financial sector to help ailing carmakers General Motors, Chrysler and Ford.

22nd – China cut interest rates by 27 basis points to 5.31%, its fifth reduction in four months.

30th – The US Treasury unveiled a US$6bn. rescue package for GMAC, the car-loan arm of General Motors, aimed at encouraging GMAC to offer funding to potential vehicle buyers.

January 2009

8th – The Bank of England reduced interest rates by a half-point to 1.5%, the lowest level since the bank was founded in 1694.

8th – Commerzbank received h10bn. (US$13.7bn.) of capital from the German government in return for a 25% stake following liquidity problems arising from its decision to purchase Dresdner Bank from insurance company Allianz.

8th – South Korea's central bank cut interest rates from 3% to a record low of 2.5%.

9th – Official figures showed that more jobs were lost in the USA in 2008 than in any year since the Second World War, with 2.6 m. axed. The jobless rate increased to 7.2% in Dec. 2008, its highest level in 16 years.

13th – China's exports fell by 2.8% in Dec. compared to the previous year, the largest decline in ten years.

13th – German chancellor Angela Merkel unveiled an economic stimulus package worth €50bn. (US$67bn.), including public investments and tax relief.

14th – The UK government guaranteed up to £20bn. of loans to small and medium-sized businesses.

14th – Shares in Europe and the USA fell sharply following the release of official figures showing a 2.7% fall in US retail sales in Dec. London's FTSE 100 closed down by over 5%, the main markets in France and Germany lost nearly 4.5% and the US Dow Jones index fell by 3%.

15th – The European Central Bank slashed interest rates by a half-point to 2%, its lowest level since Dec. 2005.

16th – The Irish government moved to nationalize Anglo Irish Bank.

16th – Reporting a fourth quarter loss of US$8.29bn., Citigroup announced plans to split into two new firms, Citicorp and Citi Holdings.

16th – Bank of America received US$20bn. of fresh US government aid and US$118bn. worth of guarantees following losses incurred in its takeover of Merrill Lynch. Merrill Lynch posted a fourth-quarter loss of US$15.3bn. while Bank of America lost US$1.7bn. in the same period.

19th – Spain became the first triple-A rated nation to have its credit rating downgraded since Japan in 2001.

19th – Denmark offered up to 100bn. kroner (US$17.6bn.) in loans to help recapitalize its banks.

20th – The French government offered its ailing car industry up to h6bn. (US$7.7bn.) in aid.

23rd – The UK economy officially entered recession after figures showed a fourth-quarter fall in GDP of 1.5% following a 0.6% drop the previous quarter.

25th – The French government provided €5bn. (US$6.5bn.) in credit guarantees to help Airbus.

26th – Dutch banking and insurance group ING estimated fourth-quarter losses of €3.3bn. (US$4.3bn.), prompting it to seek state guarantees, replace its chief executive and shed 7,000 jobs.

28th – The IMF warned that world economic growth would fall to 0.5% in 2009, its lowest level since the Second World War, and projected the UK economy would shrink by 2.8%, the worst contraction among developed nations.

28th – The International Labour Organization claimed 51 m. jobs could be lost in 2009, pushing the world unemployment rate to 7.1% compared with 6.0% at the end of 2008.

28th – Canada's Conservative government unveiled a $40bn. CDN (US$32bn.) stimulus

plan including tax cuts and infrastructure spending.

29th – New Zealand's central bank reduced interest rates by 1.5% to 3.5%.

February 2009

3rd – The Australian government announced a second stimulus package of $A42bn. (US$26.5bn.) to boost long-term growth, including one-off cash payments to low-income families and investment in infrastructure. The Reserve Bank of Australia reduced interest rates by one percentage point to 3.25%, its lowest level in 45 years.

5th – The Bank of England slashed interest rates by a half-point to a record low of 1%.

5th – Deutsche Bank unveiled a fourth-quarter loss of €4.8bn. (US$6.1bn.) and a net loss for 2008 of €3.9bn. (US$5bn.) – its first yearly loss since being restructured after the Second World War – citing 'unprecedented' operating conditions and 'weaknesses in our business model.'

9th – Barclays announced a pre-tax profit of £6.1bn. (US$9bn.) for 2008, down 14% on profits for the previous year.

9th – The French government agreed to provide Renault and Peugeot-Citroën with €3bn. (US$3.9bn.) each in preferential loans in return for maintaining jobs and sites in France. Renault Trucks, owned by Volvo, was offered a loan of €500m. (US$650 m.), suppliers €600m. (US$780 m.) and the financing arms of the two carmakers loan guarantees of up to €2bn. (US$2.6bn.).

10th – Former bosses of RBS and HBOS, two of the UK's largest financial casualties, apologized 'profoundly and unreservedly' for their banks' failure during the UK Treasury Committee's inquiry into the banking crisis.

10th – UBS declared a Swiss corporate history record loss of SFr19.7bn. (US$17bn.) for 2008 after suffering a net loss of SFr8.1bn. (US$7bn.) in the fourth quarter, including SFr3.7bn. (US$3.2bn.) in exposure to toxic assets. The bank announced it would axe a further 2,000 jobs at its investment banking arm.

12th – The Bank of Korea reduced interest rates by 50 basis points to a record low 2%.

12th – The Irish government revised its rescue plans for Allied Irish Bank and the Bank of Ireland. Each bank was to receive €3.5bn. (US$4.5bn.) and would be expected to increase lending and reduce senior executives' pay while remaining in the private sector.

12th – The Spanish economy fell into recession for the first time in 15 years, having shrunk by 1% in the fourth quarter of 2008.

17th – US President Barack Obama signed his US$787bn. economic stimulus plan after Congress approved the package.

18th – Taiwan fell into recession after its economy slumped by 8.4% in the fourth quarter. Taiwan's central bank reduced interest rates by a quarter point to 1.25%.

19th – The Bank of Japan bought 1trn. yen (US$10.7bn.) in corporate bonds and maintained a near-zero interest rate.

26th – RBS unveiled a loss of £24.1bn. (US$34.2bn.), the largest annual loss in UK corporate history, stemming from a £16.2bn. (US$23bn.) writedown of assets mainly linked to its purchase of ABN Amro. The bank also announced it would put £325bn. of toxic assets into a new government insurance scheme, while the government would inject a further £13bn. to strengthen its balance sheet.

27th – The European Bank for Reconstruction and Development (EBRD), the European Investment Bank (EIB) and the World Bank announced a €24.5bn. (US$31bn.) joint rescue package for banking sectors in Central and Eastern Europe. The two-year initiative would include equity and debt financing and policies to encourage lending, particularly to small and medium-sized firms.

March 2009

2nd – US insurance company AIG unveiled a US$61.7bn. loss in the fourth quarter of 2008, the largest in US corporate history, and received an additional US$30bn. as part of a revamped rescue package from the US government.

2nd – HSBC, Europe's largest bank, confirmed it was looking to raise £12.5bn. (US$17.7bn.) from shareholders through a rights issue after it revealed pre-tax profits for 2008 of US$9.3bn., down 62% on the previous year.

3rd – Nationalized UK bank Northern Rock confirmed it made a loss of £1.4bn. (US$2.0bn.) in 2008.

3rd – Toyota Motors, the world's largest carmaker by sales, asked for up to US$2bn. in Japanese government-backed aid.

4th – The Australian economy shrank by 0.5% in the fourth quarter of 2008.

4th – The World Bank signed a US$2bn. contingency facility to Indonesia, the largest ever loan granted to an economy not classified as in crisis. Indonesia's central bank reduced its interest rate by 50 basis points to 7.75%.

5th – The Bank of England cut interest rates from 1% to 0.5%. The Bank also announced it was to create £75bn. of new money, called quantitative easing.

9th – Iceland nationalized Straumur-Burðarás, the last of the big four banks to be taken into public ownership.

10th – Malaysia revealed a 60bn. ringgit (US$16.3bn.) stimulus package over a two year-period, amounting to 9% of GDP. The plan contained increased spending on infrastructure, guaranteed funds for businesses, equity investments to boost the stock market and tax breaks.

14th – The G20 group of rich and emerging nations pledged a 'sustained effort' to restore global growth with low interest rates and increase funds to the IMF.

16th – Serbia opened talks with the IMF over an emergency loan worth up to €2bn. (US$2.6bn.).

18th – The Bank of Japan provided up to 1,000bn. yen (US$10bn.) in subordinated loans to its commercial banks.

18th – The US Federal Reserve pledged US$1.2trn. to buy long-term government debt and mortgage-related debt.

18th – UniCredito, one of Italy's largest banks, sought h4bn. in aid from Italian and Austrian sources.

19th – The US Treasury promised up to US$5bn. to auto parts suppliers, guaranteeing payment for products shipped.

20th – The IMF revised its global forecast for 2009, with the world economy set to shrink by between 0.5% and 1%. The world's most developed economies were expected to experience the largest contractions in GDP.

23rd – The US announced a 'Public-Private Investment Programme' to buy up to US$1trn. worth of toxic assets. The US Treasury committed between US$75bn. and US$100bn. to the program, in addition to contributions from the private sector.

25th – The IMF, along with the World Bank, European Commission and other multilateral organizations, unveiled a €20bn. (US$27.1bn.) financial rescue package for Romania. The agreement stipulated Romania reduce its budget deficit to less than 3% of GDP by 2011.

25th – Italian bank Banca Popolare di Milano became the fourth bank in the country to seek funding from the government's €12bn. bank aid scheme. The bank requested €500m.

26th – Official statistics revealed that Ireland's economy shrank by 7.5% in the fourth quarter of 2008 compared to the same period the previous year, its largest contraction in decades. For the whole of 2008, the economy contracted by 2.3%, its first fall since 1983.

26th – The US economy contracted at an annualized rate of 6.3% in the fourth quarter of 2008, its fastest rate since 1982.

27th – The UK economy shrank by 1.6% in the last three months of 2008, its largest fall in GDP since 1980 and higher than the earlier 1.5% estimate.

29th – The German government pumped €60m. (US$80 m.) into Hypo Real Estate in return for an 8.7% stake.

30th – The Spanish government, with the Bank of Spain, launched a €9bn. (US$12bn.) bailout of savings bank Caja Castilla La Mancha, the country's first bank rescue in the financial crisis.

31st – The World Bank predicted the global economy would contract by 1.7% in 2009, the first decline since the Second World War. The

forecast claimed that the most developed economies would shrink by 3%, while world trade would fall by 6.8%.

April 2009

2nd – The G20 agreed to tackle the global financial crisis with fresh measures worth up to US$1.1trn. Pledges included US$750bn. made available to the IMF to help troubled economies and US$250bn. to boost global trade.

6th – Japan unveiled its latest stimulus package worth 10trn. yen (US$98.5bn.), equivalent to 2% of GDP.

7th – The Reserve Bank of Australia reduced its benchmark rate by a quarter point to 3%, its lowest level since 1960.

7th – RBS announced it would shed a further 9,000 jobs from its global operations over the next two years.

14th – Goldman Sachs reported a higher than expected pre-tax quarterly profit of US$1.8bn. The bank would also place US$5bn. worth of shares on the stock market in order to repay an emergency US$10bn. loan provided by the US government in 2008.

14th – Poland's government approached the IMF to secure a US$20.5bn. credit line to increase bank reserves and make Poland 'immune to the virus of the crisis and speculative attacks.'

14th – Fortis bank posted a loss of €20.6bn. (US$27.5bn.) for 2008 following writedowns on debt and a separation of the business.

15th – UBS unveiled a first quarter loss of SFr2bn. (US$1.75bn.) and announced it would cut 8,700 jobs by 2010 in an effort to reduce costs.

16th – China's growth rate slowed to 6.1% in the first quarter of 2009, its slowest pace since quarterly GDP data was first published in 1992. Growth was down from 6.8% in the previous quarter and 9% for the whole of 2008.

16th – Consumer prices in the USA fell by 0.4% over the year to March owing to weak energy and food prices, the first year-on-year drop since Aug. 1955.

16th – JP Morgan Chase reported a higher than expected first quarter profit of US$2.1bn. compared with net income of US$2.4bn. in the first quarter of 2008.

18th – The IMF formally agreed a US$47bn. credit line for Mexico under its new fast track scheme to help developing nations cope with the global financial crisis.

21st – UK annual inflation as measured by the Retail Prices Index (RPI) was −0.4% in March (down from zero in Feb.), the first negative figure since 1960.

21st – Sweden's central bank reduced its key interest rate by a half point to a record low of 0.5%.

22nd – UK chancellor Alistair Darling admitted the economy faced its worst year since the Second World War as he unveiled his latest Budget report. The annual budget deficit would rise sharply to £175bn. over the next two years with total government debt to reach 79% of GDP by 2013.

22nd – The IMF said global output would contract by 1.3% in 2009, a 'substantial downward revision' of its Jan. forecasts when it predicted growth of 0.5%. The UK economy was now projected to shrink by 4.1% in 2009, while Germany was set to decline by 5.6% and Japan by 6.2%.

22nd – India's central bank slashed interest rates for the sixth time in six months, reducing its key repo lending rate by a quarter point to 4.75%.

27th – National Australia Bank, Australia's largest lender, announced a 9.4% fall in cash earnings to A$2bn. (US$1.4bn.) for the Sept.–March period.

28th – Fears over a swine flu outbreak continued to have an impact on global shares – the FTSE100 closed down by 1.7%, markets in Paris and Frankfurt ended nearly 2% down, Japan's Nikkei index fell by 1.7% and Hong Kong's Hang Seng shed 1.4%.

28th – Lithuania's economy contracted by 12.6% in the first quarter of 2009 compared to the same period in 2008, the largest year-on-year fall in the EU since the start of the recession.

29th – US output contracted at an annualized rate of 6.1% in the first quarter of the year, a higher-than-expected result. The contraction was led

by a 30% decline in exports, its largest fall in 40 years.

May 2009

1st – US carmaker Chrysler filed for Chapter 11 bankruptcy protection after a group of hedge and investment funds refused to restructure the company's US$6.9bn. debt.

1st – The Reserve Bank of New Zealand reduced interest rates by 50 basis points to a record low of 2.5%. The bank governor, Alan Bollard, said he expected rates to remain at the current (or lower) level until the latter part of 2010.

4th – The European Commission forecast that the EU economy would contract by 4% in 2009, more than twice the level predicted at the beginning of the year. It claimed unemployment would now reach 10.9% in 2010.

5th – Japan offered US$100bn. of financial assistance to Asian economies affected by the global economic slowdown in a meeting of the finance ministers of the ten-member Association of South East Asian Nations.

5th – UBS confirmed it had made a SFr2bn. (US$1.75bn.) loss in the first quarter of 2009.

6th –Volkswagen and Porsche agreed to merge, relieving the sports carmaker of its debt burden.

7th – Barclays announced a pre-tax profit of £1.37bn. (US$2.07bn.) for the first three months of the year, up 15% from the previous year.

7th – Commerzbank agreed to relinquish the core of its commercial property lending business together with Eurohypo's role in public sector finance, in a deal with European competition authorities to compensate for €18.2bn. (US$24.2bn.) of state aid it received.

7th – The European Central Bank cut its main interest rate by a quarter point to a record low of 1% and also announced plans to purchase €60bn. (US$80.4bn.) of covered bonds, which are backed by mortgage or public sector loans.

7th – The Bank of England announced it would pump a further £50bn. (US$75bn.) into the UK economy in a substantial expansion of its program of government bond purchases.

8th – RBS reported a pre-tax loss of £44m. for the first quarter of 2009, compared with a profit of £479m. for the same period the previous year.

8th – Several US banks unveiled plans to raise cash a day after the US Treasury said that ten of America's 19 largest banks failed their stress tests and needed to raise a combined total of $74.6bn. Wells Fargo and Morgan Stanley planned to raise US$7.5bn. and US$3.5bn. respectively through share sales, while Bank of America planned to sell assets and raise capital to secure US$33.9bn. it needed.

13th – Franco-Belgian bank Dexia, which had been bailed out by three economies the previous year, posted a first quarter profit of €251m. (US$341 m.) compared to a loss of €3.3bn. (US$4.5bn.) in 2008.

13th – The German cabinet agreed a 'bad bank' scheme, in which banks would be able to swap their toxic debt for government-backed bonds in return for paying an annual fee.

14th – Spain suffered a fall in GDP of 1.8% in the first quarter of 2009, its largest contraction in 50 years, according to the National Statistics Institute.

14th – Crédit Agricole unveiled a net profit of €202m. (US$275 m.) in the first quarter, a 77% fall from the same period the previous year, after more than doubling its loan-loss provisions to €1.1bn.

15th – According to Eurostat economies that make up the eurozone contracted by 2.5% in the first quarter of 2009, a higher-than-forecast decline.

15th – The EBRD revealed plans to invest a record €7bn. (US$9.4bn.) in 2009 to tackle the slowdown through investments in infrastructure, energy, corporate and finance projects.

17th – Carmaker General Motors announced plans to close up to 1,100 dealerships in the USA as it battled to reduce costs and stave off bankruptcy.

19th – Inflation in the UK as measured by the Consumer Prices Index (CPI) slowed to 2.3% in April from 2.9% the previous month.

20th – Japan's GDP slid by 4% in the first quarter, its largest decline since records began in 1955.

20th – Venezuela experienced its slowest rate of growth in five years, with GDP growing by 0.3% in the first quarter of 2009 as the fall in oil prices took effect.

21st – The Office for National Statistics said public sector net borrowing in the UK rose to £8.46bn. in April compared to £1.84bn. in the same month the previous year. Concerned about its significant debt burden, Standard & Poor's downgraded the UK's credit rating from 'stable' to 'negative' for the first time since it began analyzing its public finances in 1978.

22nd – Private equity firms paid US$900 m. to rescue BankUnited, a Florida-based bank worth around US$13bn. It had been closed by federal regulators in what was the biggest US bank failure of 2009 so far.

22nd – The US Treasury provided automotive financing group GMAC with a further US$7.5bn. in state aid to help it stay in business and offer loans to potential Chrysler and GM car buyers.

22nd – UK output declined by an unrevised 1.9% in the first quarter of 2009, according to figures published by the Office for National Statistics.

26th – South Africa fell into recession for the first time since 1992 following an annualized contraction of 1.8% and 6.4% in the previous two quarters.

27th – Riksbank announced it was raising foreign currency to boost its US$22bn. currency reserves, causing a sharp fall in the Swedish krona as the central bank warned the worst of the financial crisis may not be over.

29th – India's economy grew by 5.8% in the first quarter of 2009, higher than forecast but down from 8.6% in the same quarter the previous year.

June 2009

1st – US car manufacturer General Motors filed for Chapter 11 bankruptcy protection, the biggest failure of an industrial company in US history.

2nd – Switzerland officially entered recession after the economy contracted by 0.8% in the first three months of 2009, following a decline of 0.3% in the final quarter of 2008.

3rd – Australia recorded a 0.4% rise in GDP for the first quarter compared to the same period last year, bucking international trends.

3rd – Lloyds Banking Group announced plans to cut 530 jobs and close one site in the UK by the end of 2009.

4th – Industrial and Commercial Bank of China (ICBC), the world's second largest bank by market value, unveiled plans to buy 70% of Bank of East Asia's Canadian unit as part of a move to expand overseas.

4th – The Bank of England kept interest rates unchanged at 0.5% for the third month in a row.

8th –The OECD claimed the pace of decline among its 30 member countries was slowing – the composite leading indicators index (CLI) rose 0.5 point in April.

9th – Lloyds Banking Group announced it was to shut all 164 Cheltenham & Gloucester branches, putting 1,660 jobs at risk.

9th – UK unemployment rose by 244,000 to 2.22 m. in the first three months of the year according to the Office for National Statistics (ONS), the largest quarterly rise in the jobless rate since 1981.

9th – Official figures showed that exports in Germany were 4.8% lower in April than in March and 28.7% down on the previous year, the biggest annual fall since records began in 1950.

10th – The European Central Bank provided an emergency €3bn. to the central bank in Sweden, whose banks dominate the Baltic region's financial sector.

10th – BP's annual statistical review indicated that global oil consumption fell by 0.6% in 2008, the first fall since 1993 and the largest drop since 1982.

10th – Ten of the largest US banks gained permission from the US Treasury to repay US$68bn. in government bail-out money received through the Troubled Asset Relief Programme (TARP).

11th – Figures revealed that Chinese exports fell by a record 26.4% in May from the same month the previous year.

11th – Revised GDP growth figures showed Japan contracted by 3.8% in the first quarter of 2009, less than the original estimate of 4%.

15th – The Confederation of British Industry (CBI) predicted the UK economy would contract by 3.9% in 2009 before seeing a return to growth of 0.7% in 2010.

15th – The IMF revised its growth forecast for 2010 for the USA, claiming that the economy would now grow by 0.75% compared to its forecast of 0% earlier in the year.

16th – The Bank of Japan said that the economy was no longer deteriorating, a more positive assessment than the previous month when it had stated that the economy was continuing to worsen. Nonetheless, it maintained interest rates at 0.1%.

16th – China introduced an explicit 'Buy Chinese' policy as part of its economic stimulus program, leading to fears of an increase in protectionism across the world.

17th – The US government announced a major reform of banking regulation to curb excessive risk-taking among big banks and to prevent future financial crises. President Obama described the reforms as 'the biggest shake-up of the US system of financial regulation since the 1930s.'

17th – The OECD revised its growth forecast for Italy, predicted the economy would grow by 0.4% in 2010 compared to a previously estimated contraction of 0.4%. However, it downgraded its forecast for 2009 from a 4.3% decline to 5.3%.

17th – The World Bank raised its GDP growth forecast for China to 7.2% in 2009 from a previously estimated 6.5%, citing the impact of a fiscal stimulus package.

18th – Official figures showed inflation in India had turned negative for the first time since 1977. Wholesale prices fell 1.61% in the year to 6 June.

22nd – The Japanese government looked set to provide up to 100bn. yen (US$1bn.) in state aid to Japan Airlines, the country's biggest airline, on condition that the organization's management improves.

24th – The OECD said the world economy was near the bottom of the worst recession in post-war history and predicted that the 30 most industrialized countries would shrink by 4.1% in 2009. UK output was predicted to contract by 4.3% in 2009 and experience zero-growth in 2010.

24th – The European Central Bank pumped €442.2bn. (US$628bn.) in one-year loans into the eurozone's weakened banking system in an effort to unlock credit markets and revive the region's economies.

24th – Orders for new durable goods in the USA rose unexpectedly by 1.8% in May from the previous month, going against expectations of a drop of 0.9%.

25th – The IMF said that Ireland's economy would contract by 8.5% in 2009 and warned it would experience the worst recession in the developed world and struggle to bail out its banks.

26th – New Zealand suffered a fifth straight quarterly contraction after official figures showed the economy shrank by 2.7% in the first quarter of 2009.

26th – Consumer prices in Japan fell by 1.1% in May compared to the same month the previous year, its biggest fall since records began in 1970, fuelling fears of a new bout of deflation.

26th – Spain unveiled a €9bn. (US$12.7bn.) fund aimed at saving banks suffering during the downturn.

30th – Eurozone inflation turned negative for the first time since records began in 1991, with consumer prices 0.1% lower in June than twelve months earlier.

30th – Malaysia launched economic liberalization measures aimed at attracting foreign investments, including changes to its long-standing policy of giving preferential treatment to the country's ethnic Malay majority.

July 2009

1st – Japan's Shinsei Bank and Aozora Bank merged to create the country's sixth largest bank with assets of 18trn. yen (US$186bn.).

1st – Unemployment in Ireland reached 11.9% in June, its highest level since 1996.

1st – India's exports were down 29.2% in May from the same month the previous year, the economy's eighth consecutive fall in exports.

7th – Inflation in the Philippines fell to 1.5% in June, its lowest level in 22 years.

10th – US carmaker General Motors (GM), 61% owned by the US government, emerged from its bankruptcy protection after creating a 'new GM' made up of four key brands, including Cadillac.

13th – The US deficit moved above US$1trn. for the first time in history.

14th – Inflation in the UK fell below the Bank of England's target rate of 2% for the first time since 2007. Lower food prices caused the Consumer Prices Index to drop to an annual rate of 1.8% in June, down from 2.2% in May.

14th – Singapore grew at an annualized rate of 20.4% in the second quarter, its first quarterly expansion in a year following a revised contraction of 12.7% from January to March.

14th – Goldman Sachs reported a net profit of US$3.44bn. for the second quarter of the year, higher than analysts had forecast.

15th – UK unemployment increased by a record 281,000 to 2.38 m. in the three months to May, its highest level in over ten years.

15th – Japan's central bank downgraded its economic forecast to a contraction of 3.4% from 3.1% for the 12 months to end-March 2010, but reiterated that the worst of the recession was over.

15th – Russia's economy contracted by 10.1% in the first half of 2009, its sharpest decline since the early 1990s.

16th – China's economy grew at an annualized rate of 7.9% in the second quarter, up from 6.1% between January and March, as the government upgraded the growth forecast to 8% for 2009 as a whole.

16th – JP Morgan Chase unveiled a second quarter profit of US$2.72bn., an increase of 36% on the same period the previous year.

17th – Ghana secured a US$600 m. three-year loan from the IMF and was given access to a further US$450 m. from the IMF through the special facility set up by the G20 summit to assist poor countries.

20th – Iceland announced a 270bn. kr. (US$2.1bn.) recapitalization plan for its banking system, issuing bonds to three new banks set up in 2008 following the collapse of the country's three main banks.

21st – UK government debt increased to £799bn., or 56.6% of UK GDP, its highest level since records began in 1974.

22nd – The National Institute of Economic and Social Research (NIESR) predicted UK GDP to fall by 4.3% in 2009 and UK GDP per capita to remain below its pre-recession levels until March 2014.

22nd – Morgan Stanley reported a loss of US$159bn. in the second quarter of 2009, compared to a US$698 m. profit for the same period the previous year.

23rd – Credit Suisse unveiled a 29% increase in second quarter net profits of 1.57bn. Swiss francs (US$1.48bn.).

23rd – The Asian Development Bank said growth in East Asia, excluding Japan, would double to 6% in 2010, compared to a 3% expansion in 2009.

23rd – The rate of decline of Japan's exports slowed in June, a sign that government stimulus spending around the world may be supporting demand. However, exports were still 35.7% lower than the same month the previous year.

24th – The IMF approved a 20-month Stand-By Arrangement for Sri Lanka worth US$2.6bn. to support the country's economic reform package.

24th – The UK economy contracted by 0.8% in the second quarter of 2009, much lower than the 2.4% decline in the previous quarter but above analysts' 0.3% prediction.

24th – The South Korean economy grew by 2.3% from April to June, its fastest expansion in five-and-a-half years.

28th – Deutsche Bank unveiled a net profit of €1.09bn. (US$1.56bn.) for the second quarter of 2009, a 67% increase in profits compared to the same period the previous year.

28th – BBVA, Spain's second largest bank, reported a net profit of €1.56bn. (US$2.23bn.) for the second quarter thanks to higher income from loans.

31st – Mizuho Financial Group revealed a net loss of 4.4bn. yen (US$46 m.) for the second quarter, its fourth consecutive quarterly loss.

31st – Japan's jobless rate increased by 830,000 in June to 3.48 m., its highest level in six years.

31st – Eurozone unemployment reached 9.4% (or 14.9 m. people) in June, its highest level in ten years.

August 2009

3rd – Barclays announced a pre-tax profit of £2.98bn. (US$5bn.) for the first six months of the year with an 8% increase in revenue.

3rd – HSBC saw pre-tax profits halve to £2.98bn. (US$5bn.) for the first half of 2009 compared to the same period the previous year, following the write-off of US$13.9bn. of bad debt in the USA, Europe and Asia.

3rd – World stock markets were boosted by brighter economic data – Standard & Poor's 500 index tipped beyond 1,000 for the first time since Nov. 2008, London's FTSE closed at its highest rate since Oct. 2008, the three major US indexes added over 1.25% by the end of trade after positive manufacturing survey results from July and European indexes also rose.

4th – UBS reported a loss of SFr1.4bn. (US$1.32bn.) in the second quarter, an improvement on the SFr2bn. loss made in the previous quarter.

4th – UniCredito, Italy's largest bank, unveiled better-than-expected second quarter earnings of €490m. (US$706 m.), 9.2% higher than the previous quarter.

5th – Société Générale announced a second quarter profit of €309m. (US$445 m.), 52% lower than the same period 12 months earlier.

6th – The Bank of England injected a further £50bn. into the UK economy as part of its quantitative easing program, bringing its total spending to £175bn.

6th – Commerzbank made a €763m. (US$1.1bn.) net loss in the second quarter, a small improvement on the h861m. loss registered in the previous quarter.

7th – RBS reported a pre-tax profit of £15m. for the first six months of the year.

7th – Italy's economy shrank by 0.5% in the second quarter, its fifth consecutive quarterly contraction but an improvement on the record 2.7% fall in Jan.–March.

7th – The IMF and Angola began talks on a loan to help the African country cope with the global economic slowdown.

12th – Dutch financial services group ING announced a €71m. (US$100 m.) profit in the three months to the end of June, its first profit in three quarters.

12th – Commonwealth Bank of Australia, the country's second largest bank by market capitalization, posted net earnings of A$4.72bn. (US$3.89bn.), 1% lower than the previous year owing to higher bad debt charges and reduced wealth management unit income.

12th – The UK unemployment rate increased to 7.8% in the second quarter, its highest level since 1995.

13th – France and Germany both recorded second quarter growth figures of 0.3%, bringing a year-long recession to an end. However, the Eurozone contracted by 0.1%, its fifth consecutive quarterly fall in output.

14th – Colonial BancGroup, a property lender based in Montgomery, Alabama, became the largest bank in the USA to collapse in 2009.

14th – The Nigerian Central Bank injected N400bn. (US$2.6bn.) into five banks and sacked their managers, after the regulator claimed the banks were undercapitalized and posed a risk to the entire banking system.

14th – Hong Kong posted growth of 3.3% between April and June following four consecutive quarters of contraction. Singapore also announced its emergence from recession, with annualized growth of 20.7% in the second quarter of 2009.

14th – South Africa's central bank slashed its lending rate by a half-point to a four-year low of 7%, its sixth cut since Dec. 2008.

17th – Japan's economy grew by 0.9% in the second quarter of 2009, ending a run of four consecutive quarters of negative growth.

18th – The South African economy contracted for the third quarter in a row as output fell at an annualized rate of 3% between April and June.

18th – The CPI measure of inflation in the UK remained at the same level of 1.8% in July,

although economists had forecast a decline to 1.5%.

20th – The UK's public sector net borrowing totalled £8bn. in July, the first July deficit for 13 years, as the government's overall debt reached its highest level since 1974 at 56.8% of GDP.

20th – Mexico's economy contracted by 10.3% in the second quarter owing to a decline in demand for exports and falling levels of tourism resulting from the outbreak of swine flu in April and May.

24th – Thailand posted growth of 2.3% in the second quarter of 2009 as it emerged out of recession.

26th – The Malaysian economy expanded by 4.8% in the second quarter of 2009 following two straight quarters of contraction.

27th – US GDP shrank at an annualized rate of 1% in the second quarter, lower than the 1.5% decline predicted by many economists.

27th – Credit Agricole, France's largest retail bank, announced a higher-than-expected second quarter profit of €201m. (US$286 m.).

28th – The Office for National Statistics (ONS) revised the rate of contraction in the UK economy for the second quarter to 0.7% from the original estimate of 0.8%.

28th – Unemployment in Japan hit a record high of 5.7% in July and consumer prices fell by 2.2% compared to a year earlier, its fastest recorded pace.

31st – The Eurozone's annual rate of inflation fell by 0.2%, its third consecutive monthly decline.

September 2009

1st – India's exports fell at an annualized rate of 28% in July, its tenth consecutive monthly contraction.

2nd – The *de facto* government of Honduras received US$150 m. from the IMF to boost its dollar reserves.

2nd – The OECD predicted that the recession in Iceland, marked by a large contraction in domestic demand, would be deeper than in most developed economies.

3rd – The OECD forecast the UK to be the only G7 economy to stay in recession at the end of 2009, while the eurozone and the USA would record two quarters of growth.

4th – The G20 group of nations agreed to continue fiscal stimulus until the recovery from recession was assured.

5th – The IMF sanctioned US$510 m. to Zimbabwe, its first loan to the country in a decade, to replenish the economy's dwindling foreign currency reserves.

8th – The EBRD announced it would invest a record €8bn. (US$11.6bn.) in central and eastern Europe in the course of 2009.

8th – Estonia's GDP shrank at an annualized rate of 16.1% in the second quarter of 2009, its sixth consecutive quarterly contraction. Latvia contracted by 18.7% and Lithuania by 19.5% in the same period.

8th – The gold price climbed above $1,000 per ounce for the first time since Feb. on the back of a weakening dollar and lingering concerns over the sustainability of the world economy's recovery.

9th – The FTSE 100 broke through the 5,000-point barrier for the first time since Oct. 2008.

11th – Brazil emerged from recession after it grew by 1.9% between April and June following two successive quarters of contraction.

14th – The European Commission predicted that the eurozone would grow by 0.2% in the third quarter and 0.1% in the fourth quarter, but GDP for the year would fall overall by 4%.

15th – Consumer Price Index inflation in the UK measured 1.6% in Aug., its lowest level since Jan. 2005.

15th – US Federal Reserve chairman Ben Bernanke claimed recession in the US was 'very likely over' but the economy would remain weak for some time owing to unemployment.

16th – Unemployment in the UK rose by 210,000 in the three months to July to take the total to 2.47 m., its highest level since 1995.

17th – The UK Office for National Statistics reported flat sales volumes in August compared with July, confounding analyst expectations of a 0.2% rise.

18th – The UK's public sector net borrowing totalled a record £16.1bn. in Aug., with government's overall debt standing at £804.8bn., or 57.5% of GDP.

20th – A further two US banks were closed by the country's federal regulator, taking the total number of US banks failing in 2009 to 94. Irwin Union Bank & Trust and Irwin Union Bank were shut down after their parent firm, Irwin Financial, failed to meet a Federal Deposit Insurance Corporation demand to boost their capital.

21st – The pound fell to its lowest level against the euro for five months as concerns continued about the underlying health of the British economy.

22nd – The Asian Development Bank made an upward revision of its growth forecast for India and China in 2009, with India expected to grow by 6.0% (up from an earlier forecast of 5.0%) and China by 8.2% (up from 7.0%).

23rd – The US dollar fell to a one-year low against the euro with traders switching to other currencies as signs of economic recovery emerged.

23rd – The World Bank announced it was to provide India with US$4.3bn. to fund infrastructure projects and support companies needing credit.

24th – Loss-making carrier Japan Airlines asked for a government bailout following recently announced plans to cut 6,800 jobs.

26th – Speaking at the end of the two-day G20 summit, US President Barack Obama said the world's leading nations had agreed to 'tough new measures' to prevent another global financial crisis, including regulation relating to the amount of money banks hold in reserve and a cap on pay for bankers.

29th – The Office for National Statistics revised growth figures for the UK in the second quarter from −0.7% to −0.6%.

29th – Core consumer prices in Japan fell 2.4% in Aug. year-on-year, the fourth successive month of contraction.

30th – The IMF slashed its forecast for the amount of bad debt likely to be written off globally between 2007 and 2010 from US$4.0trn. to US$3.4trn.

In Oct. 2009 US manufacturers reported that global output was growing at its fastest rate for five years. On 29 Oct. the Department of Commerce announced that the US economy was out of recession, growing by an annualized 3.5% in the third quarter. However, rising unemployment was an ongoing concern, standing at 10.2% in Oct. 2009 (its highest rate since 1983). US president Barack Obama responded to the news of the emergence from recession with caution, commenting: 'We anticipate that we are going to continue to see some job losses in the weeks and months to come.'

By the end of the third quarter of 2009, of the G7 economies only the UK remained in recession, having contracted by 0.4% in the period July–Sept.

This is an edited and updated version of the Credit Crunch Chronology that appears on The Statesman's Yearbook Online: http://www.statesmansyearbook.com/entry.html?entry=chronology_credit

See Also

- ▶ Banking Crises
- ▶ Great Depression

Credit Cycle

P. Bridel

Abstract

It was with the post-First World War attempts to integrate marginalist value and monetary theory that theorists started pondering the possible (in Hayek's words) 'incorporation of cyclical phenomena into the system of economic equilibrium theory'. Hayck's own 'intertemporal equilibrium' approach overturned the traditional view of cycles as temporary deviations from long-period equilibrium

conditions. But the publication of Keynes's *General Theory* redirected research efforts towards the determination of output at a point in time. Since the late 1960s, with the search for 'microfoundations for macroeconomics', this line of thought has been back on the theoretical agenda.

Keywords

Bank rate; Cambridge School; Capital theory; Credit cycle; Cumulative process; Forced saving; Full employment saving; Hawtrcy, R. G.; Hayek, F. A. von; Intertemporal equilibrium; Keynes, J. M.; Marginal revolution; Microfoundations; Monetary theory of interest; Natural rate and market rate of interest; Quantity theory of money; Rational expectations; Robertson, D.; Saving and investment; Temporary equilibrium

JEL Classifications

E3

Prior to Keynes's *General Theory*, the resolution of the question why, in capitalist economies, aggregate variables undergo repeated fluctuations about the trend was regarded by economists as a main challenge for the profession. What was then called business (or trade) cycle theory grew quite independently from the classical and subsequently neoclassical corpus of price theory. In fact, for all economists, a clear-cut distinction existed between the long-run forces at work in an economy – the subject of a rigorous value and distribution theory – and the more or less ad hoc explanations of the short-run oscillations around such an (equilibrium) centre of gravity. Of course, from Ricardo and Thornton down the 19th century to Overstone and Mill, money and credit played a substantial, but independent, part in these exogenous explanations of the business cycle. Along the same line, the founding fathers of marginalism (in particular Walras, Marshall and Jevons) failed to coordinate, even in a remotely satisfactory way, money and trade cycle with their then novel price theory.

Following Wickscll's and Miscs's lead, it is only with the post-First World War attempts to integrate marginalist value and monetary theory that theorists started pondering the possible 'incorporation of cyclical phenomena into the system of economic equilibrium theory' (von Hayek 1929, p. 33n.). The rediscovery of Tooke's (1844) income approach to the quantity theory of money is probably one of the earliest stepping-stones in the development of credit-cycle theories. This line of thought suggests that the explanation of money prices should start not from the quantity of money but from nominal income. Though another way of writing a Marshallian cash balance equation, Wicksell's (1898, p. 44) or Hawtrey's (1913, p. 6) emphasis on the 'aggregate of money income', on how it varies, is expanded or held, is a crucial turning-point on the road towards an analysis in terms of income, saving and investment. This shift of emphasis, together with the simultaneous progress in monetary theory proper (notably the development of a comprehensive and integrated monetary theory of interest), the 1914–1918 inflationary episode and the post-war cyclical upheavals provided in the 1920s and 1930s the right intellectual stimulus for credit-cycle theories to grow and multiply.

Explicitly or implicitly, to tackle this issue, Continental economists (for example, Mises, Cassel, Hayck, Schumpeter and Aftalion), members of the Cambridge School then dominating in England (Keynes, Robertson, Pigou, Hawtrey), Fisher and Mitchell in the United States all used the common analytical framework established jointly by Walras, Menger, Marshall and Jevons. This is made up of two basic (though familiar) propositions: on the one hand, there is an inverse relation between the volume of investment and the rate of interest (that is, a downward-sloping investment demand curve) and, on the other, despite short-run 'frictions', the interest rate is assumed to be sensitive enough to divergences between investment decisions and full employment saving.

The central theme of this argument (first expressed with great clarity in Wicksell's cumulative process) is that the market rate of interest oscillates in the short run around a natural rate of

interest determined in the long run by the supply of and the demand for capital as a stock, which, in turn, guarantees the equality between planned investment and full employment saving. Once this logic is understood, it then emerges that the entire development of interwar trade-cycle theories took place within the second proposition outlined above; namely, that, in the long run, the interest rate is assumed to be sensitive enough to divergences between investment decisions and full employment saving. Hence, since the twin concepts of an interest-elastic demand curve for investment and natural rate of interest were never called into question, the orgy of debates that took place in the 1920s and 1930s was conducted in terms of an analysis of various short-run forces which temporarily keep at bay the long-run forces of saving and investment.

These forces are, of course, of multiple nature. Of particular interest to interwar economists, and one of the essential features of business cycle, with its recurrence of upswings and downswings, is a *credit cycle,* an alternation of credit expansion and credit contraction. But it was assumed neither that an alternation of prosperity and depression would not exist in a barter economy (or in a purely specie system) nor that cycles could be viewed as functions of monetary factors only.

In fact, and thanks to their common capital theory, none of the leading interwar credit cycle theorists fell into either of these traps. Even Hawtrey who, with remarkable consistency kept claiming that business cycles are a purely monetary phenomenon, had clearly in mind a Wicksell-like cumulative process derived from Marshall's oral tradition in monetary theory. This common theoretical background and a deep interest in a then fast-developing monetary theory make similarities between credit cycle theorists.

> sufficiently pronounced to entitle us to speak of a single monetary theory [of the cycle], the votaries of which disagree on one issue only: whether bank-loan rates act primarily on 'durable capital' [Keynes, Robertson, Hayek] or via the stocks of wholesalers [Hawtrey]. (Schumpeter 1954, p. 1121)

In 1913, Hawtrey was amongst the first to provide a detailed analysis of the financial working of the cumulative process in an Anglo-Saxon environment. However, even if his theory usefully describes the ways in which money and credit behave in the cycle, the main weakness of his contribution is, of course, its almost exclusive emphasis on dealers' stocks in the course of a credit cycle. If Hawtrey does not deny altogether that a credit expansion/contraction has an influence on the volume of investment, he holds it however to be unimportant when compared with the direct influence on the wholesalers' stocks. He then logically disputes the existence of forced saving on the very ground of this availability of stocks and fails completely to link his credit cycle theory with the dominant Marshallian capital theory. Such a model led Hawtrcy not only to give Bank Rate the crucial part to play in any counter-cyclical policy but also to consider its fluctuations as the only explanation of cyclical fluctuations. To sketch British interwar depressions as almost exclusively functions of Bank Rate (itself a function of Britain's absorption of gold) is a rather bold simplification Hawtrcy was never quite ready to abandon.

If the theoretical apparatus underlying the *Treatise on Money* proceeds from the same logic, Keynes's fundamental equations introduce, however, a number of very sophisticated and new variations on the basic credit-cycle theme. In particular, causes of credit cycles are of non-monetary nature (they result from fluctuations in the rate of investment relative to the rate of saving), the influence of Bank Rate on investment is not limited 'to one particular kind of investments, namely, investments by dealers in liquid goods [stocks]' (Keynes 1930, vol. 1, p. 173), the cumulative process includes a theory of the demand for money beyond the traditional income motive (that is, an early version of liquidity preference), and, in the short run, there is no longer a direct relation between the quantity of money/credit and the price level: monetary or credit changes do not foster ipso facto a forced/abortive saving process. Despite the higher degree of sophistication shown in the *Treatise,* in a classic chapter on the modus Operandi of the Bank Rate, Keynes displays bold confidence in this mechanism to smooth any credit cycle, to fill the gap between saving and investment and to correct all temporary monetary

divergences from the long-run full employment equilibrium. However, Keynes's disaffection with the forced saving doctrine and the purely static nature of his fundamental equations drew sharp criticisms from Robertson and Hayek. Though from different standpoints, they both considered Keynes's credit cycle analysis as no more than an attempt to spell out the appropriate banking policy which could maintain a monetary equilibrium. In particular, Keynes's version of the credit cycle lacked, for the former, a proper sequential stability analysis and, for the latter, an explicit integration with capital theory.

Along lines very similar to Keynes's and, up to the late 1920s, in close cooperation with him, Robertson worked out a detailed sequential analysis of the interdependence of real and monetary magnitudes during the cycle. But clearly, for him, the cycle results from over-investment, this tendency to over-invest being a typical feature of decentralized economies stemming from the repercussions on the volume of investment of its gestation period. However, the largest part of Robertson's professional output was devoted to studying the monetary or credit symptoms of such economic fluctuations, that is, how banks may respond to an increased demand for credit during expansion.

This led Robertson to a redefinition of the concept of saving in a monetary economy and to the role of this new concept in the cycle. This approach was linked with a sequential analysis of the lagged adjustments of output to monetary flows. In the 'forced saving' debate, central to all credit cycle theories, and contrary to Hayek who considered it as the villain of the piece, Robertson saw that phenomenon as only a relatively minor component of his theory, the factors at the root to his 'credit inflation' being the *real* cause of this expansion. Dragged among others by Keynes into endless discussions in the realm of monetary and interest theory, Robertson never managed however to offer an articulate and full-blown version of his theory of industrial fluctuations. In particular, the problem of the alteration in the structure of production, a question forming the core of Hayek's cycle theory, never received more than passing comment.

Grounded of course in the Austrian tradition and Wicksell's cumulative process (first extended by Mises 1912, and Cassel 1918), the distortion of the production time structure is absolutely central to Hayek's monetary cycle theory. The divergence between 'natural' and market rates of interest is linked by Hayek to the variability in forced saving and considered as the cause of cyclical fluctuations. Hayek's 'additional credit' theory places the cause of this gap between these two rates upon newly created money. The increase in loan capital resulting from a 'trailing market rate' makes investment surpass voluntary saving: a cumulative expansion results. Such an increase in investment alters the relative prices of capital and consumer goods in favour of the former. The increased output of capital goods distorts the production time structure. At a later stage, higher factor incomes drive up the demand for consumption goods, which through increased withdrawals from bank accounts will raise the market rate of interest and, finally, make some investment unprofitable. Then, the turnabout that takes place in the cycle brings a change in the other direction in the production structure, this time in favour of consumer goods. Clearly, crises are caused by over-investment, that is, by a decline in the desire to purchase the flow of capital goods coming on the market. The reversal of the process initiated by credit inflation does take place (as in most credit cycle theories) whenever the market rate catches up with prices; and since, sooner or later, banks run up against the limits set to their lending by their reserves, this process cannot be explosive (Fisher 1911, also noticed, at least in his earliest writings, this stabilizing influence of the banking system).

Hayek's credit cycle theory thus marks a real break with what had come before. The theory of money is no longer a theory of the value of money 'in general' because relative prices may be changed by monetary influences and the Wicksellian full-employment assumption is dropped. The specific task of the trade cycle theorist is, for Hayek, to analyse short-period positions of the economy 'in successive moments of time' (1941, p. 23). The adoption of such an 'intertemporal equilibrium' approach to cycles

(conceptually not different from modern temporary equilibrium) marks not only a crucial methodological turning point, but also the swan song of credit cycle theories.

On the one hand, this new method of 'intertemporal equilibrium' heralds the abandonment of the traditional framework in which cycles (defined as short-run discquilibria) are seen as temporary deviations from long-period equilibrium conditions determined by systematic and persistent forces at work in decentralized economies. In the present case, the 'natural' rate of interest determined in the long run by the supply of and the demand for capital is no longer the norm towards which the system is tending. It is in fact a property of such an 'intertemporal equilibrium' that not only will the price of the same commodity be different at different points in time but also that the stock of capital will not yield a uniform 'natural' rate of interest on its supply-price.

On the other, the publication of Keynes's *General Theory* redirected research efforts away from this question into the problem of the determination of output at a point in time. It is only since the late 1960s, with the search for 'microfoundations for macroeconomics', and the subsequent advent of rational expectations and non-Walrasian equilibria, that this line of thought has been back on the theoretical agenda. However, given the extreme complexity of the problem and the relative crudeness of models still in their infancy, progress has so far been very modest.

See Also

► Hawtrey, Ralph George (1879–1975)

Bibliography

Fisher, I. 1911. *The purchasing power of money.* New York: Macmillan.

Hawtrey, R.G. 1913. *Good and bad trade*. London: Constable.

von Hayek, F.A.. 1929. *Monetary theory and the trade cycle*. Trans: N. Kaldor and H.M. Cromc. London: Jonathan Cape, 1933.

von Hayek, F.A. 1941. *The pure theory of capital*. London: Routledge.

Keynes, J.M. 1930. *A treatise on money,* The pure theory of money, vol. 1. As in *Collected writings*, vol. 5. London: Macmillan, 1971.

von Miscs, L. 1912. *The theory of money and credit.* Trans. H.E. Batson. London: Jonathan Cape, 1934.

Schumpeter, J.A. 1954. *History of economic analysis*. London: Oxford University Press.

Tooke, T. 1844. *An inquiry into the currency principle*. London: Longman, Brown, Green & Longmans.

Wicksell, K. 1898. *Interest and prices*. Trans. R.F. Kahn. London: Macmillan, 1936.

Credit Rating Agencies

Joel Shapiro

Abstract

Credit Rating Agencies (CRAs) have been measuring the credit risk of debt for slightly over 100 years. The industry is characterised by artificial and natural barriers to entry and an issuer-pays system. The agencies' ratings performed poorly for structured finance products and have been criticised for being an important factor in the financial crisis of 2007–2009. The critique focuses on poor modelling techniques and conflicts of interest.

Credit rating agencies (CRAs) measure the credit risk of debt for all types of investors. Their measurement of credit risk includes default probabilities and they rate both corporate and public debt. In recent years they have also expanded dramatically into structured finance investments. In general, the CRAs use hard public information that is available to all investors, and hard private and soft private information that is provided by the issuer.

CRAs serve an economic purpose: they reduce asymmetric information about issuers that investors face when making investments, thus enhancing market liquidity. They also decrease wasteful duplication of research and information production. Reputation is critical in maintaining their incentives to produce

quality ratings: short-term gains from inflating an investment's quality can be smaller than long-term losses from jaded investors.

Keywords

Conflicts of interest; Corporate bonds; Credit ratings; Fitch; Investment grade; Moodys; Securities and Exchange Commission (SEC); Standard & Poor's; Structured finance

JEL Classifications

D82; G14; G24; G28

History

Bond rating and the establishment of formal CRAs began in 1909 when John Moody began rating US railroad bonds, soon expanding to utility and industrial bonds. Poor's Publishing Company followed in 1916 and Fitch Publishing Company in 1924. The business was characterised by the investor-pays model, where investors bought reports from the CRAs containing their ratings. This changed in 1970, for two reasons. First, with the advent of the photocopier free-riding became commonplace and CRAs found it difficult to sustain their business (White 2002). Second, in 1970 Penn Central defaulted on its commercial paper obligations, creating vast mistrust among investors and a large demand by issuers for certification. The business thus changed to an issuers-pay model (Cantor and Packer 1995). In 1975, the Securities and Exchange Commission (SEC) created the Nationally Recognized Statistical Rating Organization (NRSRO) category to designate credit ratings agencies whose ratings were recognised as being valuable for investment decisions. Standard & Poor's, Moody's and Fitch were given this designation immediately, and four other firms attained it in the following 17 years. By 2000, however, mergers returned the number of NRSROs to the big three. The SEC gave out a fourth NRSRO designation in 2003 (Dominion), a fifth in 2005 (A.M. Best), and in response to congressional legislation promoting transparency and entry in 2006 gave out three more designations (White 2010). All of these new NRSROs, however, remain very small players in the bond and structured finance businesses.

Important Aspects of Industry Structure

1. Many regulatory agencies use ratings in evaluation, e.g. to determine capital requirements. Moreover, certain entities such as banks, insurance and pension funds are restricted to invest only in *investment grade* securities, i.e. BBB and above (see Cantor and Packer 1995). This creates an artificial demand for ratings. Kisgen and Strahan (2010) demonstrate that the acquisition of NRSRO status for Dominion Bond Rating Service in 2003 changed the impact of its ratings on bond yields only in situations where this status was important. Coval et al. (2009) provide evidence that Collateralized Debt Obligations (CDOs) were inaccurately priced because ratings were overly weighted by investors. Adelino (2009) finds that while initial yields on tranches below AAA for mortgage backed securities predict future credit performance the initial yields on AAA tranches had no predictive power. This is consistent with the hypothesis that investors in AAA tranches had no other information beyond the credit ratings themselves.
2. There are large *barriers to entry* in the credit rating industry: Since Congress, local governments, and regulatory agencies adopted the NRSRO designation and used it for the determination of investment grade securities (point 1), this created an 'absolute barrier to entry' (White 2002). Moreover, the need to build a reputation in order to receive business is a natural barrier to entry.
3. The fact that Moody's and S&P rate some corporate bonds which they are not paid for by issuers using public information (*unsolicited ratings*) is controversial. While the firms state that they are providing a service demanded by investors, some parties have raised the point that these ratings may be used to discipline issuers. Poon (2003) demonstrates

that unsolicited ratings tend to be lower in general, but correcting for selection does not explain all of the variation.

4. CRAs have been able to avoid *liability* for problems with ratings. Under Section 11 of the Securities Act of 1933 they were immune from misstatements. Moreover, in court they have used the argument that ratings are speech and not recommendations on how to invest (Partnoy 2002). The Dodd-Frank Financial Reform Bill passed recently exposes CRAs to liability by defining them as experts.
5. The market for corporate bond ratings is different from the market for *structured finance* ratings. Standard & Poor's and Moody's rate all corporate bonds, while the percentage that Fitch rates has been increasing. Most structured finance products receive at least two ratings, but who is rating it depends on the deal (see Ashcraft et al. 2009). The corporate bond market is established and relatively simple, and the models used are well accepted. Structured finance products are fairly new but have grown rapidly; between 1997 and 2003 global structured finance issuance grew from about $280 billion to $800 billion (Committee on the Global Finance System 2005). These products are very complex and the methods for rating structured products have been imprecise. Errors in the ratings agencies' data, assumptions and modelling have been found. Moreover, agencies are not required to perform due diligence on underlying loans and have difficulties retaining their best employees (Partnoy 2002).
6. In the structured finance market, ratings *shopping* can occur. This means that if an issuer is unhappy with a rating, it may solicit another one, either from the same CRA or from another CRA. Moreover, 'typically the rating agency is paid only if the credit rating is issued' (US SEC 2008).

Evidence on CRAs in the Corporate Bond Market

There has been a large focus on the effect of announcements on the pricing of both bonds and stocks. The main finding is the asymmetry between downgrades and upgrades: downgrades have a significant negative impact on price, but there is virtually no price change following an upgrade. The effect of ratings changes on price is complex, as the impact of ratings changes is different for firms with low ratings than for firms with high ratings. Overall, there is a clear consensus that information provided by CRAs has an effect on price (Hand et al. 1992; Hite and Warga 1997; Berger et al. 2000; Kliger and Sarig 2000; Dichev and Piotroski 2001; Jorion and Zhang 2007). These findings suggest a role for CRAs in the allocation of capital process.

In terms of accuracy, Cantor and Packer (1995) show that ratings order corresponds to default rankings. Hilscher and Wilson (2009) argue that rating agencies do a poor job at forecasting default probabilities, but capture systematic default risk.

Fitch is generally thought of as having higher ratings than Standard & Poor's and Moody's (Jewell and Livingston 1999). Becker and Milbourn (2009) finding that increased competition from Fitch's increased market share in the corporate bond market led to more issuer-friendly ratings and also less informative ratings. Bongaerts et al. (2009) however, only find a certification role for Fitch in breaking ties between Moody's and Standard & Poor's.

Structured Finance Products and the Financial Crisis of 2007–2009

Much attention has been paid to CRAs as a potential contributor to the financial crisis. The structured finance market collapsed and even 'the highest rated (AAA) mortgage-backed securities (as measured by the corresponding credit default swaps prices) fell by 70 percent between January 2007 and December 2008' (Pagano and Volpin 2009), implying that ratings were not of high quality. There is debate over whether poor quality ratings were the fault of (i) conflicts of interest, (ii) imprecise modelling, or some mixture of both.

An SEC investigation found that senior analytical managers and supervisors participated in fee discussions with issuers and the analytical staff

also discussed ratings decisions and methodology in the context of fees and market share (US SEC 2008). In addition, CRAs offer related consulting services, such as pre-rating assessments (of what a rating might be).

A few recent theoretical papers study the implications of shopping for ratings. Bolton et al. (2010) demonstrate that competition among CRAs may reduce welfare due to shopping by issuers. Faure-Grimaud et al. (2009) look at corporate governance ratings in a market with truthful CRAs and rational investors. They show that issuers may prefer to suppress their ratings if they are too noisy. They also find that competition between rating agencies can result in less information disclosure. Skreta and Veldkamp (2009) also assume that CRAs truthfully relay their information and demonstrate how noisier information creates more opportunity for shopping by issuers to take advantage of a naive clientele.

In terms of conflicts of interest, Mathis et al. (2009) find that reputation cycles may exist where a CRA builds up its reputation by relaying information accurately only to take advantage of this reputation to later inflate ratings. Bolton et al. (2010) show that conflicts of interest for CRAs may be higher when reputation costs are lower and there are more naïve investors. Bar-Isaac and Shapiro (2010) demonstrate that CRAs incentives to produce accurate ratings are likely to be countercyclical, i.e. lower in a boom than in a recession. In Pagano and Volpin (2008), CRAs have no conflicts of interest, but can choose ratings to be more or less opaque depending on what the issuer asks for. They show that opacity can enhance liquidity in the primary market, but may cause a market freeze in the secondary market.

In empirical evidence, Mathis et al. (2009) show that, controlling for economic variables, the fraction of structured finance tranches that were rated AAA has increased over the period 2000–2008. Ashcraft et al. (2009) examine subprime and Alt-A mortgage backed securities (MBS) during the period leading up to the subprime crisis and find evidence that ratings become less conservative right at the height of the MBS market peak in 2005–2007. In particular, they demonstrate that ratings quality was worse on low documentation mortgages. Griffin and Tang (2009) look at CRA adjustments to their models' predictions of credit risk in the CDO market and find that the adjustments were overwhelmingly positive, were positively related with future downgrades, and the amount adjusted increased sharply from 2003 to 2007. Benmelech and Dlugosz (2009) find that securities rated by only one agency were 6.1% more likely to be subsequently downgraded and point to shopping as the reason.

See Also

- ▶ Barriers to Entry
- ▶ Bonds
- ▶ Public Debt
- ▶ Reputation

Acknowledgments I thank Larry White for helpful comments.

Bibliography

Adelino, M. 2009. *Do investors rely only on ratings? The case of mortgage-backed securities*. Mimeo: MIT.

Ashcraft, A., P. Goldsmith-Pinkham, and J. Vickery. 2009. *MBS ratings and the mortgage credit boom*. Mimeo: Federal Reserve Bank of New York.

Bar-Isaac, H., and J. Shapiro. 2010. *Ratings quality over the business cycle*. Mimeo: NYU.

Becker, B., and T. Milbourn. 2009. *Reputation and competition: Evidence from the credit rating industry*. Mimeo: Harvard Business School.

Benmelech, E., and J. Dlugosz. 2009. *The credit rating crisis*. Mimeo: Harvard University.

Berger, A., S. Davies, and M. Flannery. 2000. Comparing market and supervisory assessments of bank performance: Who knows what when? *Journal of Money, Credit and Banking* 32: 641–667.

Bolton, P., X. Freixas, and J. Shapiro. 2010. *The credit ratings game*. Mimeo: Columbia University.

Bongaerts, D., K.J.M. Cremer, and W. Goetzmann. 2009. *Tiebreaker: Certification and multiple credit ratings*. Mimeo: Yale University.

Cantor, R., and F. Packer. 1995. The credit rating industry. *Journal of Fixed Income*: 1–26.

Committee on the Global Finance System, Bank for International Settlement. 2005. The role of ratings in structured finance: Issues and implications. http://www.bis.org/publ/cgfs23.pdf?noframes=1

Coval, J.D., J.W. Jurek, and E. Stafford. 2009. Economic catastrophe bonds. *American Economic Review* 99: 628–666.
Dichev, I., and J. Piotroski. 2001. The long-run stock returns following bond ratings changes. *Journal of Finance* 56: 173–203.
Faure-Grimaud, A., E. Peyrache, and L. Quesada. 2009. The ownership of ratings. *RAND Journal of Economics* 40(2): 234–257.
Griffin, J.M., and D.Y. Tang. 2009. *Did subjectivity play a role in CDO credit ratings?* Mimeo: UT-Austin.
Hand, J., R. Holthausen, and R. Leftwich. 1992. The effect of bond rating agency announcements on bond and stock prices. *Journal of Finance* 47(2): 733–752.
Hilscher, J., and M. Wilson. 2009. *Credit ratings and credit risk*. Mimeo: Oxford University.
Hite, G., and A. Warga. 1997. The effect of bond-rating changes on bond price performance. *Financial Analysts Journal* 53: 35–51.
Jewell, J., and M. Livingston. 1999. A comparison of bond ratings from Moody's, S&P and Fitch. *Financial Markets, Institutions and Instruments* 8: 1–45.
Jorion, P., and G. Zhang. 2007. Information effects of bond rating changes: The role of the rating prior to the announcement. *Journal of Fixed Income* 16: 45–59.
Kliger, D., and O. Sarig. 2000. The information value of bond ratings. *Journal of Finance* 55: 2879–2902.
Kisgen, D.J., and P.E. Strahan. 2010. Do regulations based on credit ratings affect a firm's cost of capital? *Review of Financial Studies* 23: 4324–4347.
Mathis, J., J. McAndrews, and J.C. Rochet. 2009. Rating the raters: Are reputation concerns powerful enough to discipline rating agencies? *Journal of Monetary Economics* 56(5): 657–674.
Pagano, M., and P. Volpin. 2008. *Securitization, transparency, and liquidity*. Mimeo: Università di Napoli Federico II and London Business School.
Pagano, M., and P. Volpin. 2009. Credit rating failures: Causes and policy options. Macroeconomic stability and financial regulation: Key issues for the G20, CEPR.
Partnoy, F. 2002. The paradox of credit ratings. In *Ratings, rating agencies and the global financial system*, ed. R.-M. Levich, G. Majnoni, and C. Reinhart. Boston: Kluwer.
Poon, W.P.H. 2003. Are unsolicited credit ratings biased downward? *Journal of Banking and Finance* 27: 593–614.
Skreta, V., and L. Veldkamp. 2009. Ratings shopping and asset complexity: A theory of ratings inflation. *Journal of Monetary Economics* 56(5): 678–695.
United States Securities and Exchange Commission. 2008. *Summary report of issues identified in the Commission Staff's Examinations of Select Credit Rating Agencies*. Washington: SEC.
White, L. 2002. The credit rating industry an industrial organization analysis. In *Ratings, rating agencies and the global financial system*, ed. R.M. Levich, G. Majnoni, and C. Reinhart. Boston: Kluwer.
White, L. 2010. The credit rating agencies: How did we get here? Where should we go? *Journal of Economic Perspectives* (forthcoming).

Credit Rationing

Charles W. Calomiris, Stanley D. Longhofer and Dwight M. Jaffee

C

Abstract

Credit rationing – a situation in which lenders are unwilling to advance additional funds to borrowers at the prevailing market interest rate – is now widely recognized as a problem arising because of information and control limitations in financial markets. This article reviews various motivations behind research on credit rationing, traces the history of theoretical efforts to explain how this phenomenon can persist in equilibrium, and reviews recent empirical research on its prevalence and effects. In the process, credit rationing is shown to be simply an extreme case of the more general problem of capital market misallocation.

Keywords

Adverse selection; Asset substitution; Asymmetric information; Availability doctrine; Banking crises; Bankruptcy; Capital market misallocation; Credit market discrimination; Credit markets in developing countries; Credit rationing; Efficient allocation; Financial intermediaries; Financial repression; Information economics; Interest rate controls; Limited liability; Liquidity shocks; Monetary policy transmission; Moral hazard; Rational expectations; Separating contracts; Usury

JEL Classifications

D8

Broadly speaking, 'credit rationing' refers to any situation in which lenders are unwilling to advance additional funds to a borrower even at a higher interest rate. In the words of Jaffee and Modigliani (1969, pp. 850–1), 'credit rationing

[is] a situation in which the demand for commercial loans exceeds the supply of these loans at the commercial loan rate quoted by the banks'. Key to this definition is that changes in the interest rate cannot be used to clear excess demand for loans in the market. In essence, this definition treats credit rationing as a supply side phenomenon, with the lender's supply function becoming perfectly price inelastic at some point.

If the projects that are being funded by the loan are not scalable, however, then a distinction must be made between a situation in which a lender eventually restricts the size of loan it will provide to any individual borrower and one in which 'rationed' borrowers are denied credit altogether. This phenomenon arises in circumstances in which lending is not scalable. Stiglitz and Weiss (1981, pp. 394–5) therefore define credit rationing as follows:

> We reserve the term credit rationing for circumstances in which either (a) among loan applicants who appear to be identical some receive a loan and others do not, and the rejected applicants would not receive a loan even if they offered to pay a higher interest rate; or (b) there are identifiable groups of individuals in the population who, with a given supply of credit, are unable to obtain loans at any interest rate, even though with a larger supply of credit, they would.

According to this definition, lenders fully fund some borrowers but deny loans to others despite the fact that the latter are identical in the lender's eyes to those who receive loans.

Thus, there are two working definitions of credit rationing in the literature. The first focuses on situations in which increases in the interest rate cannot clear excess demand in the loan market, whether this excess demand reflects a single borrower (who would like a larger loan amount) or many. Under this definition, rationing would exist if every potential borrower received a loan but a smaller one than that desired at the equilibrium interest rate. The second definition – the Stiglitz–Weiss definition – restricts its attention to situations in which some borrowers are completely rationed out of the market, even though they would be willing to pay an interest rate higher than that prevailing in the market.

Both of these definitions focus on the supply side of the market. One could argue, however, that it is useful to think of non-price rationing as any phenomenon that limits the amount of funding used by firms such that firms are not able to use the price mechanism to successfully bid for additional funds, whether this is caused by supply-side constraints (as under the narrow definitions of credit rationing described above) or by other distortions in credit markets (related, for example, to regulation). This would allow a broader definition of 'credit rationing' in which regulatory constraints, rather than just informational problems, lead to non-price allocations of credit.

Why Care About Credit Rationing?

Early interest in credit rationing was driven in part by questions about the role that credit rationing might play in transmitting the macroeconomic effects of monetary policy, which was related to research on the so-called 'availability doctrine' in the 1950s and 60s (Scott 1957). To the extent that monetary policy operates through a 'credit channel' (in which contractionary policy affects the economy through a decline in the supply of funds available for banks to lend), and to the extent that changes in the terms of lending include not only changes in loan pricing but also changes in the quantities of credit available to borrowers, credit rationing may play an important role in the transmission of monetary policy's effects on the economy (Blinder and Stiglitz 1983).

In addition to the cyclical effects of rationing in credit markets related to monetary policy, development economists, especially Ronald McKinnon (1973), argued that a different credit rationing problem is more relevant for the long-term growth prospects of developing countries. High inflation, high zero-interest reserve requirements, government-mandated loan allocations to favoured borrowers, and interest rate ceilings on loans or deposits in developing economies (a combination which McKinnon termed 'financial repression') subjected many developing countries' banking systems to an extreme form of regulation-induced credit rationing. High

reserves, high inflation, and interest ceilings on deposits meant that banks were rationed in the deposit market, and thus had few funds to lend, while lending mandates and loan interest-rate ceilings meant that what funds were available to lend were often rationed by restrictions on who could bid for those funds.

Additionally, George Akerlof (1970), in his path-breaking article on the role of adverse selection in preventing market development, drew attention at an early date to the possible effects of information problems in retarding the development of lending markets, particularly in developing countries. In an ideal world, in the absence of any government policies limiting beneficial lending, all borrowers with positive net present value projects would be able to obtain outside funding (whether through debt or equity instruments, or bank or non-bank sources of funds). But Akerlof showed that, if markets were unable to distinguish good risks from bad ones, lending might not be feasible. The failure to develop institutions capable of producing credible information about borrowers and using that information to screen applicants could, according to Akerlof, play an important role in financial underdevelopment.

Many development economists have come to recognize that the failure to properly allocate funds in the loan market – a broad phenomenon, within which credit rationing is a special and extreme case – can be an especially important potential impediment to growth in developing countries because of the relative absence of institutions in those countries that allow effective screening of borrowers (to mitigate adverse selection) or ongoing monitoring of borrowers' actions (to mitigate moral hazard).

An additional motivation for an interest in credit rationing comes from the literature on bank fragility. Credit rationing can also apply to the market in which financial intermediaries raise their funds. Financial institutions go to great pains to attract and maintain deposits through (a) the structure of their contracts (which typically afford withdrawal options to depositors), (b) their long-term relationships with market monitors who track their progress, and (c) their established reputations for good management. But sometimes the market suddenly decides to ration credit to a particular bank or to the whole banking system; and when this happens the affected banks find it hard to attract and maintain deposits at any price. Thus, the literature on 'bank runs' as an historical phenomenon can be thought of as a literature on credit rationing in the markets in which financial institutions raise their funds. Depositors that decide to participate in a bank run ration credit to their bank in the sense that the decision to withdraw is a quantity, not a price, decision. They are simply unwilling to leave their money in the bank.

Finally, much of the current research on discrimination in credit markets is driven by evidence that black and Hispanic minority loan applicants are denied more frequently than comparable whites (for example, Munnell et al. 1996; Cavalluzzo and Cavalluzzo 1998; Cavalluzzo and Wolken 2005). Of course, this begs the question of why borrowers are denied loans in the first place, rather than simply priced according to their risk. In other words, understanding why there are differences in denial rates across groups necessarily entails exploring why rationing (loan denial) occurs.

The Development of Credit Rationing Theory

Early Views on Credit Rationing

The earliest discussions of credit rationing viewed it as a non-equilibrium phenomenon, arising either because of exogenous interest rate rigidities (for example, interest rate ceilings or usury laws) or because of a lack of competition in the loan market (Scott 1957). Soon authors made a distinction between temporary credit rationing, in which market interest rates are slow to adjust to exogenous shocks such as changes in the lender's cost of funds or borrower demand, and 'equilibrium' credit rationing, which persists after the market has fully adjusted to these shocks. Clearly the more interesting and difficult to explain phenomenon is equilibrium credit rationing.

Hodgman (1960) was the first to try to explain how credit rationing can persist in a rational, equilibrium framework. In this model, lenders

evaluate potential borrowers on the basis of the loan's expected return–expected loss ratio. In addition, it is assumes that there is a maximum repayment that the borrower can credibly promise, which effectively limits how much the lender will offer the borrower regardless of the interest rate: eventually the expected losses become too great relative to the expected return. This model was much debated in the ensuing years. In particular, Miller (1962) argued that Hodgman's analysis could be made consistent with rational expectations between the borrower and lender by incorporating bankruptcy costs that would be incurred by the lender upon the borrower's default. The real significance of the Hodgman article, however, was that it established as an important theoretical goal the objective of explaining how credit rationing could persist as an equilibrium phenomenon.

Freimer and Gordon (1965) resolved many of the issues regarding the structure of the Hodgman and Miller models by showing that credit rationing can occur with a risk-neutral lender if the borrower has a fixed-sized funding need. But this was done assuming an exogenous interest rate. Jaffee and Modigliani (1969) completed the picture by endogenizing the equilibrium interest rate by modelling both the supply and demand sides of the market. Credit rationing in their model, however, is the direct result of an exogenous assumption that borrowers within a given group must be charged the same interest rate, even though the lender can distinguish differences among them.

This early work was important in that it firmly established the idea that credit rationing could be a persistent equilibrium phenomenon. Ultimately, however, the solutions proposed relied on very restrictive assumptions about agent preferences or the contracts they could employ. More satisfactory explanations of credit rationing had to wait for the information economics revolution of the 1970s.

Modern Credit Rationing Theory

Akerlof's (1970) pioneering article on adverse selection was motivated in part by the desire to explain extreme cases of credit rationing (the absence of a credit market), but Jaffee and Russell (1976) provide the first explicit asymmetric information rationale for credit rationing in the general sense. In their model, lenders cannot distinguish *ex ante* between high-and low-quality borrowers (that is, those who will repay their loans and those who will default). Contracts are written to determine the size of the loan offered and the interest rate. As in the Rothschild and Stiglitz (1976) insurance framework, low-quality borrowers must accept the contract that is preferred by the high-quality borrowers, lest they be identified as the deadbeats they are. Although a market-clearing interest rate/loan amount combination does exist, high-quality borrowers prefer a contract that entails a slightly lower interest rate with a reduced loan amount. As a result, the pooling outcome entails credit rationing. The primary problem with this model is that the 'equilibrium' is not stable, in that unsustainable separating contracts dominate the pooling outcome.

In 1981, Joseph Stiglitz and Andrew Weiss published what has become the canonical model of credit rationing, because it was the first model that fully endogenized contract choices with a stable, rationing equilibrium. In the Stiglitz–Weiss framework, credit rationing occurs because the lender's expected return is not monotonically increasing in the interest rate. Instead, adverse selection or moral hazard problems eventually cause the lender's expected return to decline as the interest rate rises.

In the adverse selection version of the model, borrowers and lenders are both risk neutral. Borrowers are characterized by their projects, which are assumed to have the same expected returns but differ from one another in their risk. Specifically, borrower projects differ on the basis of mean-preserving spreads (Rothschild and Stiglitz 1970). These projects are also assumed to require a fixed investment (that is, they are indivisible) and borrowers have a fixed amount of internal equity that they can invest in the project. Limited liability upon default means that the lender's payoff is a concave function of the project's return, while the borrower's profit function is convex.

These assumptions imply that, at any given interest rate, a subset of the least risky borrowers

will drop out of the market, choosing instead to forgo their projects. In essence, the borrower's limited liability means that he reaps all of the project's gain (beyond the cost of debt service) when its return is high, but loses his collateral (his paid-in capital invested in the project, if any) only when the project's return is low. For low-risk projects, however, the potential upside gains are small. If those low-risk borrowers are pooled with high-risk borrowers, they will face higher than warranted interest rates. Low-risk borrowers will increasingly withdraw from the market as interest rates rise; as rates rise, borrowers with low-risk projects are better off withdrawing from the market and simply consuming their endowments rather than agreeing to invest and pay a high interest rate. As a result, increases in the interest rate cause more and more good borrowers to drop out of the market, lowering the average creditworthiness of the lender's remaining applicant pool. The size of the adverse selection premium faced by low-risk borrowers (the amount of interest low-risk borrowers have to pay in excess of what their project risks warrant) becomes larger with each interest rate rise because the interest rate must compensate for the default risk of an ever-worsening pool of borrowers.

Thus, increases in the interest rate affect lender returns in two ways. The first is the direct effect that a higher interest rate raises the lender's return (for a given pool of borrowers). Rising interest rates, however, also have the indirect effect of lowering the average quality of the lender's applicant pool, thereby lowering the lender's expected return from any given loan. Eventually, this secondary, adverse selection effect may outweigh the first interest rate effect, causing lender profits to decline as the interest rate rises.

Once the non-monotonicity of the lender's return in the interest rate is established, the possibility of credit rationing follows immediately. Profit-maximizing lenders will never voluntarily choose to raise the interest rate beyond where the adverse selection effect dominates. If excess demand exists in the market at this rate, credit rationing will be the equilibrium.

Paradoxically, in this model the very best credit risks do not seek funding because they do not find it worthwhile. This may seem odd, but it is important to remember that these borrowers are not rationed. Instead, they voluntarily drop out of the market because the cost of being pooled with higher-risk borrowers is too great. The rationed borrowers are the higher-risk borrowers who stay in the market and request funding.

Alternatively, Stigliz and Weiss show how changes in the interest rate may also affect the borrower's choice of project, so that moral hazard in project choice (sometimes referred to as 'asset substitution' in the finance literature) can be another reason that the lender's expected return is non-monotonic in the interest rate. Suppose that the borrower is able to choose among projects with different risk profiles. If, at a given interest rate, the borrower is indifferent between two projects, Stiglitz and Weiss show that an increase in the interest rate will cause the borrower to prefer the project that has the higher probability of default. Of course, the lender prefers the safer project. Thus (with slightly more restrictive distributional assumptions than in the adverse selection case), increases in the interest rate once again can eventually lower the lender's expected return, leading to credit rationing.

Models of credit rationing need not posit rationing for all borrowers. Realistically, some borrowers (certain firms for which information control problems are particularly acute) may be subject to rationing while other borrowers are not. Borrowers not subject to rationing may be able to avoid rationing because their prospects are more observable, or because their behaviour is more controllable.

Bank Runs as Credit Rationing

The theoretical literature on credit rationing in the deposit market (bank runs) has some features that distinguish it from the literature on credit rationing in the loan market. The ultimate causes of deposit market rationing can be similar to, or very different from, the causes of loan market rationing. As discussed above, loan market rationing can reflect either information and incentive problems in the loan market or exogenous

regulations. In the case of the deposit market, rationing can result either from incentive and information problems relating to the depositor–bank relationship or from exogenous liquidity needs of depositors.

With respect to the former, under some circumstances a bank run may reflect a loss of confidence in the market value of the bank's asset portfolio and changes in bank behaviour that attend such a loss. If the value of the portfolio falls sufficiently, and if the information and incentive problems are sufficiently severe, the perceived risk of losses in the bank can prompt depositors to ask for their money back because depositors have reason to be risk-intolerant (that is, to be unwilling to leave their money in a bank that has too high a level of risk). An example of such a model is Calomiris and Kahn (1991). Here the depositor withdraws funds in bad states of the world because doing so is necessary to prevent the banker from abusing his control over the bank's portfolio.

An alternative cause of credit rationing in the deposit market is a shock to the liquidity needs of depositors, which forces depositors to demand their funds from their banks irrespective of the portfolio performance of the banks. Diamond and Dybvig (1983) is an example of a model of this phenomenon.

Bank depositor runs are but one specific example of how financial intermediaries may be credit rationed due to creditor risk intolerance and/or liquidity shocks. During the 1998 Russian financial crisis, for example, it was widely reported that many emerging market hedge funds dumped their holdings of risky securities of all kinds in a scramble to reduce their risks and thus re-establish the high-quality credit ratings needed to retain their debtors. Intermediaries were also scrambling to accumulate liquidity, as many of their claimants needed to withdraw funds to meet other obligations related to the financial market upheaval.

The Limits of Credit Rationing

Credit rationing as a problem of information and control (as it was modelled by Jaffee and Russell 1976 and Stiglitz and Weiss 1981) is properly seen as an extreme case of the more general phenomenon of capital market misallocation, which includes cases where capital is misallocated (due to adverse selection and moral hazard) without any rationing occurring. It is important to recognize that, from the standpoint of either cyclical concerns about the transmission of monetary policy or developmental concerns about the efficiency of the allocation of capital, the important phenomenon is not rationing per se but rather the extent to which the market fails to allocate resources efficiently. Even a market that never suffers from credit rationing can be highly inefficient in its allocation of capital. In that sense, credit rationing may be somewhat beside the point. Indeed, the corporate finance literature is full of examples of models of market imperfections involving moral hazard and adverse selection in which credit is misallocated, and in which positive net present-value projects are not funded or negative net present-value projects *are* funded.

In some cases, firms may even be priced out of the market for funds entirely, so that they avoid funding profitable investments. For example, Jensen and Meckling (1976) show that the potential for asset substitution at the expense of creditors can make it much more costly for firms to access debt markets. Indeed, asset substitution can make it prohibitively expensive to issue debt. Note that this is not a case of credit rationing as defined by Stiglitz and Weiss, since suppliers are not refusing credit. Rather, the high asset substitution premium that firms would be charged if they sought credit can result in a decision by the firm not to fund a positive net present-value investment. Similarly, Myers and Majluf (1984) show that because of adverse selection problems – which are particularly acute in the public equity market – some firms may decide to avoid issuing equity to fund a positive net present-value investment. Here, again, a firm is not being rationed by suppliers, but is unwilling to seek financing because of its prohibitive pricing.

As the literature on capital market misallocations and credit rationing developed in the late 1970s and early 1980s, critics pointed out some limiting circumstances in which capital markets

did not have a tendency to underfund positive net present-value projects. For example, both adverse selection and moral hazard problems can be overcome by sufficient collateral. By placing collateral at risk a firm could signal its high quality, or commit itself not to abuse creditors by undertaking excessive risk (see Bester 1985). Of course, collateral is not always available, nor is it costless to place collateral at risk. In the case of a limited liability enterprise, the firm's net worth limits its available collateral. Firms that can finance themselves from internal funds and limited amounts of low-risk debt can avoid the adverse selection and moral hazard costs associated with external finance, but young, growing firms tend to be in need of substantial amounts of external finance, far in excess of their accumulated net worth. If borrowers use all of their available 'collateral', then, on the margin, collateral cannot mitigate adverse selection or moral hazard problems.

In the consumer context, it is also important to recognize that the moral hazard and adverse selection problems that arise in corporate lending may differ in importance across the various areas of consumer lending. For example, moral hazard may be limited in the context of mortgage lending where actions destructive to the lender's interest are likely to harm the homeowner as well (consider inadequate protection against the risk of fire, for example). Furthermore, the modern use of credit scores and loan-to-value ratios may make mortgage lenders more knowledgeable about an applicant's true credit risk than the applicant himself, particularly if that applicant has significant equity invested in the house and lacks experience in the credit market (Calomiris et al. 1994). Under such circumstances, the implications of adverse selection models (which depend on the superiority of the information of the borrower about his type) may be irrelevant, or even reversed. On the other hand, in the context of uncollateralized credit card borrowing based only on past credit records, unobservably high-risk borrowers (those who know that they are about to have major medical costs, lose their job, or become divorced) may have strong incentives to borrow, implying the possibility for severe adverse selection.

How Is Credit Rationing Measured Empirically?

Although credit rationing is a widely discussed phenomenon, there is a surprising paucity of evidence confirming its existence. The key problem is that, while the concept of a credit-rationed borrower is easy to understand in theory, under each of the various models of credit rationing discussed above it is extremely difficult to measure 'excess demand' of individual borrowers or the similitude of borrowers' creditworthiness.

Indirect Methods

Jaffee and Modigliani (1969) attempt to infer the presence of credit rationing by measuring the proportion of new commercial loans originated at the prevailing prime rate and/or with very large loan sizes. The intuition they use is that prime and/or large borrowers have the lowest risk and are therefore the least likely to be rationed. As a result, a larger proportion of loans will go to these low-risk borrowers when credit rationing is severe. Jaffee and Modigliani use this proxy to see how market factors affect the prevalence of credit rationing. Of particular interest is their result that increases in the average commercial loan rate are associated with higher levels of rationing, which seems to confirm the appropriateness of their proxy for credit rationing.

Other authors have attempted to measure whether commercial loan rates are 'sticky' in response to changes in open-market interest rates. The idea here is that in most credit rationing models there is an implicit cap above which lenders will ration credit. As open-market rates rise, this cap is more likely to become binding, meaning that commercial loan rates will not fully respond to changes in open-market rates. Following this approach, a number of authors, including Goldfeld (1966) and Jaffee (1971), have found that commercial loan rates are, in fact, slow to adjust to changes in open-market rates, and offer this as evidence in support of credit rationing.

Berger and Udell (1992), however, provide convincing evidence that, although commercial-loan rate stickiness does occur, it does so in a fashion that is inconsistent with information-

based credit rationing models. In particular, they find that nearly half of the observed loan rate stickiness occurs for loans made to borrowers who are exploiting a previously contracted bank loan commitment. Such borrowers are precluded from rationing by contract. Furthermore, they show that the fraction of loans made under commitment actually decreases during times of credit market tightness, exactly the opposite of what one would expect should credit rationing be an important phenomenon.

Direct Methods

Other authors have attempted to directly measure credit rationing using survey data to identify 'rationed' borrowers. For example, Cox and Jappelli (1990) and Chakravarty and Scott (1999) use data from the Survey of Consumer Finances (SCF) in which households are directly asked whether they recently have been denied credit or been unable to obtain as much credit as they requested. Although these articles purport to measure how some outside factor affects the likelihood of being rationed, it is not clear that borrowers who self-report being denied credit have, in fact, been 'rationed' in the Stiglitz–Weiss meaning of the term. After all, their denial of credit could simply reflect a failure to properly select into the right risk class in order to be approved, or the fact that the borrower was simply uncreditworthy at any interest rate.

With regard to business lending, Cressy (1996) uses a sample of new businesses that opened accounts with a major British bank to ascertain whether credit rationing affects the likelihood of business survival. He concludes that firms self-select for finance based on the entrepreneur's human capital, implying that no credit rationing is occurring.

One strand of the empirical literature on credit rationing, broadly defined, focuses on whether differential mortgage loan denial rates between white and minority borrowers constitutes evidence of discrimination (a much cited reference is Munnell et al. 1996; Ross and Yinger 2002, provide an excellent review of this literature). Although the discrimination literature does not specifically focus on the question of whether borrowers are credit rationed, any conclusion that one group is denied loans at a greater rate than others after creditworthiness is controlled for would imply that a form of credit rationing is occurring. This 'rationing', however, is distinct from that in Stiglitz–Weiss because the borrowers are not observably identical, and the underlying cause of 'rationing' is either lender preferences (Becker 1971) or some form of statistical discrimination (Calomiris et al. 1994; Longhofer and Peters 2005).

Evidence on 'Intermediary Rationing'

In contrast to the limited evidence of traditional borrower credit rationing, there is a significant body of evidence supporting the idea that financial institutions are rationed by their depositors. In recent years, a large literature has developed examining the determinants of deposit withdrawal from individual banks, and a parallel literature has developed on systemic banking panics. These articles find that in circumstances where the condition of banks is perceived to have deteriorated, depositors withdraw funds rather than simply demand a higher interest rate on deposits (Calomiris and Mason 2003; Calomiris and Wilson 2004). The links between bank characteristics and deposit withdrawals observed in these and other similar studies suggest that deposit rationing is related to information and incentive problems, rather than just liquidity shocks to depositors, although such shocks may still play a role.

Final Thoughts

It is worth noting that improvements in underwriting processes may have dramatically altered the practical impact of credit rationing in recent years. The use of risk-based pricing in consumer lending, including credit card loans and mortgages, has become widespread, reflecting the increased ability of lenders to

distinguish between borrowers with different risk profiles (see, for example, Edelberg 2003; Chomsisengphet and Pennington-Cross 2006). The same is true for commercial credit markets, in which instruments such as junk bonds, senior-subordinated securitization issues, and the like serve to provide financial market access to broader classes of instruments, borrowers and risks. As a result, 'sorting' among borrowers overall has increased, and today there is likely much less diversity in pools of 'observably identical' borrowers than there was when Stiglitz and Weiss first developed their model. While this suggests that in some markets credit rationing is a very different and perhaps less important phenomenon today than it once was, an important potential role remains for credit rationing, particularly as it pertains to financial allocations in emerging markets, the pricing of particularly opaque segments of the lending markets of developed economies, and the ways in which financial institutions may be rationed in response to shocks to their portfolios.

See Also

- ▶ Akerlof, George Arthur (Born 1940)
- ▶ Banking Crises
- ▶ Microcredit
- ▶ Stiglitz, Joseph E. (Born 1943)

Bibliography

Akerlof, G.A. 1970. The market for 'lemons': quality uncertainty and the market mechanism. *Quarterly Journal of Economics* 84: 488–500.

Becker, G.S. 1971. *The economics of discrimination*. 2nd ed. Chicago: University of Chicago Press.

Berger, A.N., and G.F. Udell. 1992. Some evidence on the empirical significance of credit rationing. *Journal of Political Economy* 100: 1047–1077.

Bester, H. 1985. Screening vs. rationing in credit markets with imperfect information. *American Economic Review* 75: 850–855.

Blinder, A.S., and J.E. Stiglitz. 1983. Money, credit constraints, and economic activity. *American Economic Review* 73: 297–302.

Calomiris, C.W., and C.M. Kahn. 1991. The role of demandable debt in structuring optimal banking arrangements. *American Economic Review* 81: 497–513.

Calomiris, C.W., and J.R. Mason. 2003. Fundamentals, panics, and bank distress during the depression. *American Economic Review* 93: 1615–1647.

Calomiris, C.W., and B. Wilson. 2004. Bank capital and portfolio management: The 1930s 'capital crunch' and the scramble to shed risk. *Journal of Business* 77: 421–455.

Calomiris, C.W., C.M. Kahn, and S.D. Longhofer. 1994. Housing finance intervention and private incentives: Helping minorities and the poor. *Journal of Money, Credit and Banking* 26: 634–674.

Cavalluzzo, K.S., and L.C. Cavalluzzo. 1998. Market structure and discrimination: The case of small businesses. *Journal of Money, Credit, and Banking* 30: 771–792.

Cavalluzzo, K., and J. Wolken. 2005. Small business loan turndowns, personal wealth, and discrimination. *Journal of Business* 78: 2153–2177.

Chakravarty, S., and J.S. Scott. 1999. Relationships and rationing in consumer loans. *Journal of Business* 72: 523–544.

Chomsisengphet, S., and A. Pennington-Cross. 2006. The evolution of the subprime mortgage market. *Federal Reserve Bank of St. Louis Review* 88 (1): 31–56.

Cox, D., and T. Jappelli. 1990. Credit rationing and private transfers: Evidence from survey data. *Review of Economic Statistics* 72: 445–454.

Cressy, R. 1996. Are business startups debt-rationed? *Economic Journal* 106: 1253–1270.

Diamond, D.W., and P.H. Dybvig. 1983. Bank runs, deposit insurance, and liquidity. *Journal of Political Economy* 91: 401–419.

Edelberg, W. 2003. Risk-based pricing of interest rates in household loan markets. FEDS Working Paper No. 2003–62.

Ferguson, M.F., and S.R. Peters. 2000. Is lending discrimination always costly? *Journal of Real Estate Finance and Economics* 21: 23–44.

Freimer, M., and M.J. Gordon. 1965. Why bankers ration credit. *Quarterly Journal of Economics* 79: 397–416.

Goldfeld, S.M. 1966. *Commercial bank behavior and economic activity: A structural study of monetary policy in the postwar United States*. Amsterdam: North-Holland.

Hodgman, D.R. 1960. Credit risk and credit rationing. *Quarterly Journal of Economics* 74: 258–278.

Jaffee, D.M. 1971. *Credit rationing and the commercial loan market*. New York: Wiley.

Jaffee, D.M., and F. Modigliani. 1969. A theory and test of credit rationing. *American Economic Review* 59: 850–872.

Jaffee, D.M., and T. Russell. 1976. Imperfect information, uncertainty, and credit rationing. *Quarterly Journal of Economics* 90: 651–666.

Jensen, M.C., and W.H. Meckling. 1976. Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics* 3: 305–360.

Longhofer, S.D., and S.R. Peters. 2005. Self-selection and discrimination in credit markets. *Real Estate Economics* 33: 237–268.
McKinnon, R.I. 1973. *Money and capital in economic development*. Washington, DC: Brookings Institution.
Miller, M.H. 1962. Credit risk and credit rationing: Further comments. *Quarterly Journal of Economics* 76: 480–488.
Munnell, A.H., G.M. Tootell, L.E. Browne, and J. McEneaney. 1996. Mortgage lending in Boston: Interpreting HMDA data. *American Economic Review* 86: 25–53.
Myers, S.C., and N.S. Majluf. 1984. Corporate financing and investment decisions when firms have information that investors do not have. *Journal of Financial Economics* 13: 187–221.
Ross, S.L., and J. Yinger. 2002. *The color of credit: Mortgage discrimination, research methodology, and fair-lending enforcement*. Cambridge, MA: MIT Press.
Rothschild, M., and J.E. Stiglitz. 1970. Increasing risk I: A definition. *Journal of Economic Theory* 2: 225–243.
Rothschild, M., and J.E. Stiglitz. 1976. Equilibrium in competitive insurance markets: An essay on the economics of imperfect information. *Quarterly Journal of Economics* 90: 630–649.
Scott, I.O. 1957. The availability doctrine: Theoretical underpinnings. *Review of Economic Studies* 25: 41–48.
Stiglitz, J.E., and A. Weiss. 1981. Credit rationing in markets with imperfect information. *American Economic Review* 71: 393–410.

Crime and Punishment

Isaac Ehrlich

'Economics of Crime' revives an old tradition in economic thought in its reliance on the unifying power of economic analysis to explain human behaviour and resource allocation choices both within and outside the conventional market place. Classical economists such as Beccaria, Paley, and Bentham devoted considerable attention to the explanation of crime in rational economic terms, and to the formulation of optimal rules for punishing offenders, based on utilitarian principles. Motivated, in part, by the rapid growth of reported offences in recent decades, economists have regained interest in the issue. Several studies in the 1960s, notably the seminal work by Becker (1966), have inspired the development of the 'economic approach to crime'.

The essence of the approach lies in the assumption that offenders respond to incentives, both positive and negative, and that the volume of actual offences in the population is therefore influenced by the allocation of private and public resources to law enforcement and other means of crime prevention. For this approach to provide a useful approximation of the complicated reality of crime, it is not necessary that all those who commit specific offences respond to incentives, (nor is the degree of individual responsiveness prejudged); it is sufficient that a significant number of potential offenders so behave on the margin. By the same token, the theory does not preclude a priori any category of crime, or any class of incentives. Indeed, economists have applied this approach to a myriad of illegitimate activities, from tax evasion and violations of minimum wage laws to auto-theft, skyjacking, and murder.

Theory

In Becker's analysis the equilibrium volume of crime was produced through the interaction between offenders and the law enforcement authority, and the focus was on propositions concerning the socially optimal probability, severity, and type of criminal sanction. Later work centred on a more complete formulation of the components of the system, especially the supply of offences, the production of law enforcement activities, and the criteria for optimal law enforcement. Attempts have also been made to expand the notion and scope of the 'market' for illegitimate activities by expounding the roles played by offenders (supply), consumers and potential victims (private demand), and enforcement and prevention (government intervention), and by augmenting the relevant market equilibrium analysis.

Supply

The offender's choice is generally modelled to involve an optimal allocation of time among competing legitimate and illegitimate activities which differ in the mix of their uncertain pecuniary and

non-pecuniary consequences, and offenders are presumed to act as expected-utility maximizers. The basic opportunities affecting choice are identified as the (perceived) probabilities of apprehension, conviction, and punishment, and the marginal penalties imposed ('deterrence variables'); the deterrence variables associated with related crimes; the marginal returns on competing illegal and legal activities and the risk of unemployment; and initial wealth. Entry into a specific criminal activity is shown to be related inversely to its own deterrence variables, and directly to the differential return it provides. Moreover, a one per cent increase in the probability of apprehension is shown to generate a larger deterrent effect than corresponding increases in the conditional probabilities of conviction given apprehension, and specific punishments given conviction (see Ehrlich 1975). Essentially due to conflicting income and substitution effects, some results for active offenders are more ambiguous: a strong preference for risk may reverse the deterrent effect of sanctions (Ehrlich 1973) and the results are even less conclusive if one assumes (as do Block and Heineke 1975) that the length of time spent in crime, not just the moral obstacle to entering it, generates disutility. The results become less ambiguous at the aggregate level, however, as one allows for non-homogeneity of offenders due to differences in personal opportunities or preferences for crime: a more severe sanction can reduce the crime rate by deterring the entry of potential offenders even if it has little effect on actual ones.

Demand

The incentives operating on offenders often originate with, and are partially controlled by, consumers and potential victims. Transactions in illicit drugs and stolen goods, for example, are patronized by consumers who generate a direct or derived demand for the underlying offences (cf. Vandaele 1978). But even for crimes that inflict pure harm on victims there exists an indirect (negative) demand, which is derived from a positive demand for safety. By their choice of optimal self-protective efforts through use of locks, safes, and alarms, or selective avoidance of travel, potential victims influence the marginal returns to offenders, and thus the implicit 'demand' for crime. And since optimal self-protection generally increases with the perceived risk of victimization (the crime rate), private protection and public enforcement will be interdependent.

Public Intervention

Whereas crime is an external diseconomy and crime control measures are largely a public good, collective action is needed to augment individual self-protection. Public intervention typically aims to 'tax' illegal returns through the threat of punishment, or to 'regulate' offenders via incapacitation and rehabilitation programme. All control measures are costly. Therefore, the 'optimum' volume of offences cannot be nil, but must be set at a level where the marginal cost of each measure of enforcement or prevention equals its marginal benefit.

To assess the relevant net benefits, however, one must adopt a criterion for public choice. Becker (1966) and Stigler (1970) each chose maximization of a concept of 'social income' as the relevant criterion, requiring the minimization of the sum of social damages from offences and the cost of law enforcement activities. This approach can lead to powerful propositions regarding the optimal magnitudes of probability and severity of punishments for different crimes and different offenders, or, alternatively, the optimal level and mix of expenditures on police, courts, and corrections. It reaffirms the proposition that, in equilibrium, the deterrent effect of the optimal probability of apprehension will exceed that of the conditional probabilities of conviction and of specific punishments, and it makes a strong case for the superiority of monetary fines as a deterring sanction. Different criteria for public choice, however, yield different implications regarding the optimal mix of probability and severity of punishment, as is the case when the social welfare function is expanded to include concern for the distributional consequences of law enforcement and other concepts of justice in addition to aggregate income (see Polinsky and Shavell 1979; Ehrlich 1982). Furthermore, a positive analysis of enforcement must address the behaviour of the

separate agencies constituting the enforcement system and the constraints of the political market. Studies which focus on the production of and demand for specific agencies, such as police and courts (see, e.g., Landes 1971), have often adopted decision rules which deviate from the social welfare maximizing criterion.

Market Equilibrium

A general equilibrium analysis of the market for offences involving the joint determination of the volume of offences and the net returns from crime in a system of interrelated markets is still at an embryonic stage. One important implication of the market model already developed is that the efficacy of deterring sanctions cannot be assessed merely by reference to the elasticity of the aggregate supply of offences, but depends on the elasticity of the private demand schedule as well. Likewise, the efficacy of rehabilitation and incapacitation programmes cannot be inferred solely from knowledge of their impact on individual offenders. It depends crucially on the elasticities of the market supply and demand schedules, as these determine the extent to which successfully rehabilitated offenders will be replaced by others responding to the prospect of higher net returns (see Ehrlich 1981; van den Haag 1975). A market setting has also been applied by economists to analyse various aspects of organized crime.

Empirical Analyses

Largely due to the paucity of theoretically relevant data, little has been done thus far to implement a comprehensive market model of illegitimate activity (but see Vandaele 1978). In particular, few studies have sought to estimate the private demand for self-protection as part of a complete market system (see Bartel 1975; Clotfelter 1977). Many researchers have attempted, however, to implement a simultaneous equation model of crime and law enforcement activity consisting, typically, of three sets of basic structural equations (see Ehrlich 1973): supply-of-offences functions linking the rate of offences with deterrence variables and other measurable incentives; production functions of law enforcement activity linking conditional probabilities of arrest, conviction, and punishment with resource inputs and other determinants of productivity; and demand-for-enforcement functions linking resource spending with determinants of public intervention. The bulk of the econometric work concerns the first two structural relationships. (For surveys see Palmer 1977; Andreano and Siegfried 1980; Pyle 1983.)

The econometric applications have been hampered by a number of methodological problems. For example, FBI crime reports are known to understate true crime rates, and related errors of measurement in estimated punishment risks may expose parameter estimation to biases and spurious correlations. The inherent simultaneity in the data requires systematic use of identification restrictions to assure consistent estimation of structural parameters. In testing offenders' responsiveness to incentives, estimates of the deterrent effect of imprisonment must be distinguished from those of its incapacitative effect. Efficient functional forms of structural equations must be selected systematically. And then there is the ubiquitous possibility that results would be biased by 'missing variables' (including links to markets for illicit drugs or handguns). While these problems have been recognized from the outset, not all studies have attempted to resolve them by applying relevant statistical remedies.

Most studies of specific offences report similar findings: probability and length of punishment are generally found to be inversely related to crime rates, and the estimated elasticities of the latter with respect to the conditional risk of apprehension are often found to exceed those with respect to the conditional risks of conviction and punishment. Crime rates are often found to be directly related to measures of income inequality and community wealth (presumably due to the link between affluence and criminal opportunities). Estimates of unemployment effects are somewhat ambiguous, however, depending, in part, on whether they are derived from time-series or cross-section data (see the survey by Freeman

1983), and such is the case also with demographic variables. This pattern of results is derived from studies using aggregate data from different countries and locations, FBI as well as Victimization Survey statistics, and even individual crime data. There also is some evidence that police output measures are weakly responsive to additional resource inputs, although studies differ in their definitions of output and in their specification of the relevant production functions.

Not all research, however, is consistent with the deterrence hypothesis (e.g. Forst 1976; but see its critique by Wadycki and Balkin 1979). Also, criticism has been raised as to the validity of the estimated deterrent effects on grounds of potential biases due to errors of measurement and the identification restrictions used (see Blumstein et al. 1978). Critics have argued that the apparent deterrent effects may mask a deterrent effect of crime on punishment variables. These issues are clearly debatable (see Ehrlich and Mark 1977).

The applicability of the economic approach to the crime of murder, and whether the death penalty constitutes a specific deterrent have raised greater controversy. The centre of debate has been the study by Ehrlich (1975) in which the approach was found to be not inconsistent with time-series evidence (see Blumstein et al. 1978; Ehrlich and Mark 1977). The controversy has generated additional empirical research, some inconsistent with the deterrence hypothesis (e.g., Passell 1975; Forst 1977; Avio 1979; Hoenack and Weiler 1980) and some quite corroborative (e.g. Ehrlich 1977; Wolpin 1978; Phillips and Ray 1982; Layson 1983, 1985).

It is early to assess the degree to which the various econometric studies on crime have produced accurate estimates of critical behavioural relationships. Some studies attempting to test the theory have not, in fact, taken sufficient account of it. Both theory and econometric design, however, must be further developed to account for missing elements of the general market model, thereby facilitating the substantive identification of structural equations and, indeed, the explanation of observed crime variations. While a consensus seems to emerge among researchers regarding the potential power of the economic approach in studying both the illegal sector of the economy and its interaction with the legal economy, future progress will greatly depend on better data.

See Also

- ▶ Family
- ▶ Law and Economics

References

Andreano, R., and J.J. Siegfried. 1980. *The economics of crime*. Cambridge, MA: Schenkman.

Avio, K.L. 1979. Capital punishment in Canada: A time-series analysis of the deterrent hypothesis. *Canadian Journal of Economics* 12: 647–676.

Bartel, A.P. 1975. An analysis of firm demand for protection against crime. *Journal of Legal Studies* 4(2): 433–478.

Becker, G.S. 1966. Crime and punishment: An economic approach. *Journal of Political Economy* 76(2): 169–217.

Becker, G.S., and W.M. Landes (eds.). 1974. *Essays in the economics of crime and punishment*. New York: Columbia University Press.

Block, M.K., and J.M. Heineke. 1975. A labor theoretic analysis of the criminal choice. *American Economic Review* 65(3): 314–325.

Blumstein, A., J. Cohen, and D. Nagin (eds.). 1978. *Deterrence and incapacitation: Estimating the effects of criminal sanctions on crime rates*. Washington, DC: National Academy of Science.

Carr-Hill, R.A., and N.H. Stern. 1979. *Crime. The police and criminal statistics*. London: Academic Press.

Clotfelter, C.T. 1977. Public services, private substitutes, and the demand for protection against crime. *American Economic Review* 67(5): 867–877.

Ehrlich, I. 1973. Participation in illegitimate activities: Theoretical and empirical investigation. *Journal of Political Economy* 81(3): 521–565. Reprinted with supplements in Becker and Landes (1974).

Ehrlich, I. 1975. The deterrent effect of capital punishment: A question of life and death. *American Economic Review* 65(3): 397–417.

Ehrlich, I. 1977. Capital punishment and deterrence: Some further thoughts and additional evidence. *Journal of Political Economy* 85(4): 741–788.

Ehrlich, I. 1981. On the usefulness of controlling individuals: An economic analysis of rehabilitation, incapacitation and deterrence. *American Economic Review* 71(3): 307–322.

Ehrlich, I. 1982. The optimum enforcement of laws and the concept of justice: A positive analysis. *International Review of Law and Economics* 2(1): 3–27.

Ehrlich, I., and R. Mark. 1977. Fear of deterrence. *Journal of Legal Studies* 6: 293–316.
Fleisher, B.M. 1966. *The economics of delinquency*. Chicago: Quadrangle.
Forst, B.E. 1976. Participation in illegitimate activities: Further empirical findings. *Policy Analysis* 2(3): 477–492.
Forst, B.E. 1977. The deterrent effect of capital punishment: A cross-state analysis of the 1960s. *Minnesota Law Review* 61(5): 743–767.
Freeman, R.B. 1983. Crime and unemployment. In *Crime and public policy*, ed. J.Q. Wilson. San Francisco: ICS.
Heineke, J.M. (ed.). 1978. *Economic models of criminal behavior*. Amsterdam: North-Holland.
Hoenack, S.A., and W.C. Weiler. 1980. A structural model of murder behavior. *American Economic Review* 70(3): 327–341.
Landes, W.M. 1971. An economic analysis of the courts. *Journal of Law and Economics* 14(1): 61–107.
Layson, S. 1983. Homicide and deterrence: Another view of the Canadian time-series evidence. *Canadian Journal of Economics* 16(1): 52–73.
Layson, S. 1985. Homicide and deterrence: A reexamination of the United States time-series evidence. *Southern Journal of Economics* 52(1): 68–89.
Palmer, J. 1977. Economic analyses of the deterrent effect of punishment: A review. *Journal of Research in Crime and Delinquency* 14(1): 4–21.
Passell, P. 1975. The deterrent effect of the death penalty: Statistical test. *Stanford Law Review* 28(1): 61–80.
Phillips, L. 1981. The criminal justice system: Its technology and inefficiencies. *Journal of Legal Studies* 10(2): 363–380.
Phillips, L., and S.C. Ray. 1982. Evidence on the identification and causality dispute about the death penalty. In *Applied time series analysis*, ed. O.D. Anderson and M.R. Perryman. Amsterdam: North-Holland.
Polinsky, A.M., and S. Shavell. 1979. The optimal tradeoff between the probability and magnitude of fines. *American Economic Review* 69(5): 880–891.
Pyle, D.J. 1983. *The economics of crime and law enforcement*. London: Macmillan.
Stigler, G.J. 1970. The optimum enforcement of laws. *Journal of Political Economy* 78(3): 526–535.
Tullock, G. 1967. The welfare costs of tariffs, monopolies, and theft. *Western Economic Review* 5(3): 224–232.
Van den Haag, E. 1975. *Punishing criminals*. New York: Basic Books.
Vandaele, W. 1978. An econometric model of auto theft in the United States. In ed. J.M. Heineke.
Wadycki, W.J., and S. Balkin. 1979. Participation in illegitimate activities: Forst's model revisited. *Journal of Behavioral Economics* 8(2): 151–163.
Witte, A.D. 1980. Estimating the economic model of crime with individual data. *Quarterly Journal of Economics* 94(1): 57–84.
Wolpin, K. 1978. Capital punishment and homicide in England: A summary of results. *American Economic Review: Papers and Proceedings* 68(2): 422–427.

Crime and the City

Yves Zenou

Abstract

Crime is unevenly distributed across space and tends to be concentrated in poor areas. Recent theoretical advances show that social interactions and peer effects can explain this pattern because of contagion effects and social multipliers. An individual is more likely to commit crime if his or her peers commit crime than if they do not. Recent empirical findings suggest that, indeed, social interactions and networks are key to understand criminal behaviour in cities.

Keywords

Becker, G.; Black–white wage differences; Bonacich centrality measure; Contagion effects; Cost–benefit analysis; Crime and the city; Crime rates; Crime, economic theory of; Social multiplier effects; Social networks

JEL Classifications

R29; A14; K42; R14

Crime is defined as an act committed in violation of a law forbidding it and for which punishment is imposed upon conviction. Crime is, however, not evenly distributed across space as it tends to be concentrated in specific areas where people are generally poor and uneducated. In both the United States and Europe, the typical urban pattern is that large cities have higher crime rates than smaller cities, and poor, largely minority neighbourhoods experience higher crime rates than more affluent white neighbourhoods (Raphael and Sills 2005). According to the United Nations Interregional Crime and Justice Research Institute, (see Alvazzi del Frate 1997), the percentage of population who are victims of burglary in urban areas with more than 100,000 inhabitants over a five-year period

(between 1992 and 1996) is: 16 for Western Europe, 24 for North America, 20 for South America, 18 for Eastern Europe, 13 for Asia and 38 for Africa. Another typical pattern common to both the United States and Europe is that ethnic minorities are overrepresented in criminal activities. In the United States, the proportion of 20–29-year-old black men directly in trouble with the law (in jail or prison or on probation or parole) reached 23 per cent in 1989 (Freeman 1999). There is, however, one notable difference. Since the mid-1980s, crime has declined in the United States but increased in Europe, especially in large urban areas (Blumstein and Wallman 2000).

Theories

In the standard crime model (Becker 1968), each individual has to implement a cost–benefit analysis in order to choose between becoming a criminal and participating in the labour market. The cost is the severity of punishment, which obviously depends on the probability of being arrested. The benefit consists in the proceeds from crime. If crime is localized, then criminals will trade off a lower probability of being arrested (since, in some areas, a host of criminals are active and the number of policemen is not sufficient) against lower proceeds from crime (more criminals also imply less booty). In this context, Sah (1991) examines the influence of the social environment on individuals' perceptions of the probability of arrest. Indeed, people develop their ideas about the relative benefits and costs of crime based on the observations they make every day. If a person lives in an area with a high crime rate, and particularly if the criminals are seen to be relatively successful, then that person is more likely to engage in criminal activity. The main result of this paper is that individuals in some areas tend to commit more crime than the Beckerian model would predict because of the gap between the perceived and the real cost of committing crime, which leads to a lower sense of impunity based on the information provided by their criminal friends.

Another approach (Verdier and Zenou 2004) proposes that distance to jobs plays a role in crime behaviour and provides a unified explanation for why blacks commit more crime, are located in poorer neighbourhoods and receive lower wages than whites. The mechanism is as follows. If everybody believes that blacks are more prone to crime than whites, even if there is no basis for this, then blacks are offered lower wages and, as a result, locate further away from jobs. Because distant residence implies more tiredness and higher commuting costs, the black–white wage gap is widened further. Blacks have thus a lower opportunity cost of committing crime (lower outside option) and become indeed more criminal than whites. The loop is closed and the beliefs are self-fulfilling.

Whereas the standard Beckerian approach focuses on individual behaviour, Glaeser et al. (1996) stress the role of peers and social interactions in criminal activities, especially in urban areas because of the high variance in crime rates. Two types of individuals are assumed: those who, as in the standard model, base their crime decision on a cost–benefit analysis, and those who only imitate their neighbours. Because of these social interactions, the benefits from crime are greater than in the Beckerian model. Moreover, if these interactions are localized (as is usually the case), then it becomes easy to explain very high levels of crime in some areas of the city. Indeed, if there are already a lot of criminals in a particular location, then crime becomes 'contagious' by spreading like a virus and amplifies the number of criminals in this location. There are social *multiplier* effects through a feedback loop: negative social behaviour such as crime leads to more negative social behaviour.

Calvó-Armengol and Zenou (2004), and Ballester et al. (2004) propose a model along these lines but represent social interactions in terms of a social network of criminal friends. People in a network not only imitate but also influence each other. Here, the cost of committing crime is reduced thanks to the network of friends. Indeed, delinquents learn from other criminals belonging to the same network how to commit

crime in a more efficient way by sharing the know-how about the 'technology' of crime. They show that the influence of peers on the individual's criminal activity depends on his or her position in the network, and each agent's criminal effort is proportional to his or her Bonacich centrality measure (see Bonacich 1987). For a given network, the Bonacich network centrality counts, for each agent, the total number of direct and indirect paths of any length in the network stemming from this agent. Such paths are weighted by a geometrically decaying factor (with path length). In other words, the 'location' of each individual in a network of friends, as measured by the Bonacich centrality measure, is a key determinant of his or her criminal activity.

As a result, in a spatial or social context, an efficient policy aiming at reducing crime would not be, as in the Beckerian model, to increase at random the cost of committing crime, but rather to target criminals according to their location in the urban or social space. Ballester et al. (2004) propose a policy that consists in finding and getting rid of the key player, that is, the criminal who, once removed, leads to the highest aggregate crime reduction. They show that the key player is not necessarily the most active criminal (that is, the one with the highest Bonacich centrality). Indeed, removing a criminal from a network has both a direct and an indirect effect. The direct effect is that fewer criminals contribute to the aggregate crime level. The indirect effect is that the network topology is modified, and the remaining criminals adopt different crime efforts. The key player is the one with the highest overall effect.

Empirical Studies

One of the first tests of the Becker model was undertaken by Ehrlich (1973), who used as explanatory variables the imprisonment rate and the average sentence for the crime in question. More recently, the focus has been on urban or social problems because this is particularly fruitful for understanding personal and property crime as opposed to white-collar crime. Cullen and Levitt (1999), using data for 137 US cities from 1976 to 1993, explore the relationship between crime and urban flight (that is, the flight of the white population from city centres to suburbs). They find that each additional reported crime in city centre is associated with a net decline of about one resident. Causality runs from rising crime rates to city depopulation. Pursuing this area of research, Glaeser and Sacerdote (1999) provide three reasons for higher crime rates in big cities. They report that 27 per cent of the difference between urban and rural crime rates in the United States is due to higher pecuniary benefits for crime in cities, 20 per cent to a lower probability of arrest and recognition in cities, and the remaining 45–60 per cent to the observable characteristics of individuals. This last number can be explained by a positive covariance across agents' decisions about crime, so that the variance of crime rate is higher than the variance predicted by local conditions. This implies that social interactions should matter, especially in cities.

Case and Katz (1991) were among the first to investigate this last issue. Using data from the 1989 NBER survey of youths living in low-income Boston neighbourhoods, they find that the behaviours of neighbourhood peers appear to substantially affect youth behaviours in a manner suggestive of contagion models of neighbourhood effects. The direct effect of moving a youth with given family and personal characteristics to a neighbourhood where 10 per cent more of the youths are involved in crime than in his or her initial neighbourhood is to raise the probability the youth will become involved in crime by 2.3 per cent.

Glaeser et al. (1996) find that, across crimes, crime committed by younger people has higher degrees of social interaction, while, across cities, for serious crimes in general and for larceny and auto theft in particular, the degree of social interactions is larger in those communities where families are less intact, that is, have more female-headed households. Ludwig et al. (2001) and Kling et al. (2005) explore this last result by using data from the Moving to Opportunity (MTO) experiment that assigned a total of 638 families from high-poverty Baltimore

neighbourhoods into three 'treatment groups': (a) Experimental group families receive housing subsidies, counselling and search assistance to move to private-market housing in low-poverty census tracts; (b) Section 8-only comparison group families receive private-market housing subsidies with no programme constraints on relocation choices; and (c) a Control group receives no special assistance under MTO. They show that relocating families from high- to low-poverty neighbourhoods reduces juvenile arrests for violent offences by 30–50 per cent of the arrest rate for control groups. This also suggests very strong social interactions in crime behaviours.

Using a very detailed data-set of friendship networks in the United States from the National Longitudinal Survey of Adolescent Health (AddHealth), Calvó-Armengol et al. (2005) test the main results of Ballester et al. (2004). Contrary to the standard approach, here peer effects are conceived not as an average intra-group externality that affects identically all the members of a given group, but as a collection of dyadic bilateral relationships, which constitutes a social network. The position and thus the centrality of each individual are thus crucial to understand criminal behaviour. Calvó-Armengol et al. (2005) show that, after observable individual characteristics and unobservable network specific factors are controlled for, the individual's position in a network (as measured by his or her Bonacich centrality) is a key determinant of his or her level of criminal activity. A standard deviation increase in the Bonacich centrality increases the level of individual delinquency by 45 per cent of one standard deviation.

See Also

- Law, Economic Analysis of
- Neighbours and Neighbourhoods
- Racial Profiling
- Residential Segregation
- Social Interactions (Theory)
- Social Multipliers
- Social Networks in Labour Markets
- Spatial Mismatch Hypothesis
- Urban Economics

Bibliography

Alvazzi del Frate, A. 1997. *Preventing crime: Citizens' experience across the world* (Issues and reports no. 9). Rome: United Nations Interregional Crime and Justice Research Institute.

Ballester, C., A. Calvó-Armengol, and Y. Zenou. 2004. *Who's who in crime networks. Wanted: The key player* (Discussion paper no. 4421). London: CEPR.

Becker, G. 1968. Crime and punishment: An economic approach. *Journal of Political Economy* 76: 169–217.

Blumstein, A., and J. Wallman. 2000. *The crime drop in America*. New York: Cambridge University Press.

Bonacich, P. 1987. Power and centrality: A family of measures. *American Journal of Sociology* 92: 1170–1182.

Calvó-Armengol, A., and Y. Zenou. 2004. Social networks and crime decisions: The role of social structure in facilitating delinquent behavior. *International Economic Review* 45: 939–958.

Calvó-Armengol, A., E. Patacchini, and Y. Zenou. 2005. *Peer effects and social networks in education and crime* (Discussion paper no. 5244). London: CEPR.

Case, A., and L. Katz. 1991. *The company you keep: The effects of family and neighborhood on disadvantaged youths* (Working paper no. 3705). Cambridge, MA: NBER.

Cullen, J., and S. Levitt. 1999. Crime, urban flight, and the consequences for cities. *Review of Economics and Statistics* 81: 159–169.

Ehrlich, I. 1973. Participation in illegitimate activities: A theoretical and empirical investigation. *Journal of Political Economy* 81: 521–565.

Freeman, R. 1999. The economics of crime. In *Handbook of Labor Economics*, ed. O. Ashenfelter and D. Card. Amsterdam: North-Holland.

Glaeser, E., and B. Sacerdote. 1999. Why is there more crime in cities? *Journal of Political Economy* 107: S225–S258.

Glaeser, E., B. Sacerdote, and J. Scheinkman. 1996. Crime and social interactions. *Quarterly Journal of Economics* 111: 508–548.

Kling, J., J. Ludwig, and L. Katz. 2005. Neighborhood effects on crime for female and male youth: Evidence from a randomized housing voucher experiment. *Quarterly Journal of Economics* 120: 87–130.

Ludwig, J., G. Duncan, and P. Hirschfield. 2001. Urban poverty and juvenile crime: Evidence from a randomized housing-mobility experiment. *Quarterly Journal of Economics* 116: 655–679.

Raphael, S., and M. Sills. 2005. Urban crime in the United States. In *A companion to urban economics*, ed. R. Arnott and D. McMillen. Boston: Blackwell.

Sah, R. 1991. Social osmosis and patternsof crime. *Journal of Political Economy* 99: 1272–1295.

Verdier, T., and Y. Zenou. 2004. Racial beliefs, location and the causes of crime. *International Economic Review* 45: 731–760.

C

Crises

P. Kenway

The term 'crisis' as used in economics is principally associated with Marx. While other writers use the term, Marx attempted rigorously to theorize crises as they occur in capitalism. It is therefore his work which will be discussed here.

In one sense, what Marx meant by an economic crisis accords perfectly well with the common use of the term: for example, it would be quite appropriate to use it to describe the liquidation of a company due to bankruptcy or a major financial disruption, involving the collapse of a number of banks. Marx however used the term 'crisis' rather more precisely, applying it to any situation where the process of renewal and expansion of capital was interrupted. Thus, for example, overproduction by one sector of the economy would cause a crisis, whether restricted to that one sector alone, or not. The term also includes the most general crises, affecting all branches of the economy and many national economies simultaneously.

For Marx, long periods of economic decline or stagnation were not 'crises'. Neither should it be thought that by *the* crisis is meant solely the final demise of capitalism. For crises were (and are) a normal and frequent feature of capitalism, and they represent not only a breakdown in the process of capital accumulation, but also the means through which capital reorganizes itself for a fresh burst of accumulation.

Two important points must be made about Marx's theory of crises. The first is that Marx identified the forces which give rise to the possibility of crisis within the process of capitalist production itself. While not disputing that economic crises could also arise as a result of disturbances from outside the economic sphere (such as natural disasters), these were not Marx's concern. Marx attempted to show that crises could be generated 'internally' by capitalism. The second point is to emphasize that there is a distinction within the theory between the analysis of the features of capitalism which give rise to the possibility of crisis, and the analysis of those conditions which turn this latent possibility into reality. Although the 'theory of the possibility of crisis' grows over into the consideration of crises proper, it inevitably precedes it and lays the foundation for this analysis.

Most analyses of the actual content of crises begin with the circuit of capital, M–C–M. The purpose of theory of the possibility of crisis is to show why that form, M–C–M, contains the *potential* for crisis. It is that theory which will be discussed here.

Capitalist production is the production of commodities. To show that crises were intrinsic to capitalism, Marx had therefore to develop the theory of the possibility of crisis from his analysis of the commodity.

A commodity, Marx observed, is a product produced for exchange. It is not produced to meet the needs of the person who produces it. The commodity has two sides to it, its use-value (or usefulness) which is entirely dependent on its physical properties, and its value, the magnitude of which is measured by the amount of socially necessary labour time required for its production. As it is produced for exchange, it has to pass through a series of distinct forms: firstly as 'commodity' then as money and then again as 'commodity'. This commodity circuit is usually depicted as C–M–C.

It is worth explaining this in a little more detail to avoid any ambiguity. Suppose that I manufacture an item for sale. At this stage, my commodity is in its natural or 'commodity' form. Suppose now that I succeed in selling it. My commodity now takes the form of money. It is still a commodity (money is a commodity) but it now takes the form of money where previously it took a physical form. If I now use this money to make a purchase, my commodity has now once more reverted to a natural, 'commodity' form. C–M–C refers to the phases through which the one commodity has to pass, though its circuit is of course intertwined with the circuits of other commodities. In accordance with common sense, the first phase (C–M) is the sale and the second (M–C), the purchase.

A number of observations may now be made. Since the commodity is produced for sale, it must undergo the metamorphosis from 'commodity' to money. Whether it succeeds in this depends on conditions which are external to the commodity, conditions which may or may not prevail. The fact that it must attempt this transformation, the success of which depends upon conditions external to the commodity, is what creates 'the germ of the possibility of crisis' (Marx 1861, p. 507). The possibility of crisis arises from the fact that the commodity may fail to complete this metamorphosis: it may fail to be sold.

It may seem that Marx was doing no more than state the obvious: a commodity must be sold. Such an assessment would be wrong for two reasons. It should be remembered that it is a result derived from his analysis of the commodity, not merely an assertion. Secondly, it is significant that those who deny that crises are an inevitable feature of capitalist production, do so essentially by ignoring or assuming away the very characteristics which Marx's analysis uncovered.

To illustrate this, it is worth looking at how Marx challenged Ricardo's denial of the possibility of general overproduction. Ricardo's position was that: 'Productions are always bought by productions, or by services; money is only the medium by which the exchange is effected' (Ricardo 1821, pp. 291–2). To this, Marx replied:

> Here ... the exchange of commodities is transformed into mere barter of products, of simple use-values. This is a return not only to the time before capitalist production, but even to the time before there was simple commodity production: and the most complicated phenomenon of capitalist production – the world market crisis – is flatly denied by denying the first condition of capitalist production, namely that the product must be a commodity and therefore express itself as money and undergo the process of metamorphosis. (Marx 1861, p. 501)

But if the possibility of crisis lies firstly in the simple metamorphosis of the commodity, in the commodity circuit C–M–C, it is far from fully developed. 'For the development of this possibility into reality', Marx observed, 'a whole series of conditions is required which do not yet even exist from the standpoint of the simple circulation of commodities' (Marx 1867, p. 209). Thus the theory of the possibility of crisis must be extended to take account of the implications of the circuit of capital.

Although the circulation of commodities is the starting point of capital, the circuit of capital is a dramatic transformation of that followed by the commodity. Instead of C–M–C, the capital circuit is M–C–M (Money–'Commodity'–Money). In the capital circuit, capital, as money, is firstly used to buy commodities (means of production, raw materials and labour-power). These are then put to use to produce items for sale which are then sold, if possible, at a profit. With this sale, capital has once more returned to the money form.

> It is worth noting that money plays a quite different role in C–M–C, compared with M–C–M. In the circulation of the commodity, money acts merely as money, as medium of circulation, whereas 'money which describes the latter course in its movement is transformed into capital, becomes capital, and from the point of view of its function, is capital'. (Marx 1867, p. 248)

Two more points of contrast between M–C–M and C–M–C should be mentioned. Firstly, the goal of the simple circulation of the commodity is the acquisition of further commodities for their use-value: the goal is consumption. In contrast, the driving force of the circulation of capital, its determining purpose, is exchange value (Marx 1867, p. 250). Secondly, although both C–M–C and M–C–M contain a sale phase and a purchase phase, the order of the two phases is inverted. In C–M–C, it is selling in order to buy. In M–C–M, it is buying in order to sell.

This inversion has a direct bearing on the development of the possibility of crisis. For obviously, if the circuit is broken, it will be during the sale phase. This creates a problem even under the simple circulation of commodities but its impact is likely to be limited. Once the circuit becomes a capital circuit, a failure to sell has more far-reaching consequences, because it means that the very purpose of production has been thwarted.

Marx illustrated this in his discussion on money as a means of payment. Essentially, a chain of mutual financial obligations develops: should the cloth fail to be sold, then many capitalists will be affected, not just the cloth merchant.

The weaver will not be paid; he in turn will be unable to pay the spinner; neither will be able to pay the machine manufacturer and he in turn will be unable to pay the suppliers of iron, timber and coal. 'This is nothing other than the possibility of crisis described when dealing with money as a means of payment; but here – in capitalist production – we can already see the connection between the mutual claims and obligations, the sales and purchases, through which the possibility can develop into actuality' (Marx 1861, p. 512).

Ricardo's denial of the possibility of general overproduction is now worth another look. His main argument was this:

> No man produces, but with a view to consume or sell, and he never sells but with an intention to purchase some other commodity, which may be immediately useful to him, or which may contribute to future production. By producing, then, he necessarily becomes either the consumer of his own goods, or the purchaser and consumer of the goods of some other person. It is not to be supposed that he should, for any length of time be ill-informed of the commodities which he can most advantageously produce, to attain the object which he has in view, namely, the possession of other goods; and therefore, it is not probable that he will continuously produce a commodity for which there is no demand. (Ricardo 1821, p. 290)

Marx found fault with this on three counts. Firstly, in saying that a man may produce in order to consume, Ricardo was again overlooking the fact that commodities are produced to be sold, and not to meet the needs of the producer. It is true that where production is for the direct satisfaction of the producer, there are no crises. But such a situation is not even simple commodity production, let alone capitalist production (Marx 1861, p. 502).

Marx's second criticism goes to the very heart of the matter:

> A man who has produced does not have the choice of selling or not selling. He must sell. In the crisis there arises the very situation in which he cannot sell or can only sell below the cost price or must even sell at a positive loss. What difference does it make to him or us that he has produced in order to sell? The very question we want to solve is what has thwarted that good intention of his? (Marx 1861, p. 503)

Finally, 'no man sells but with an intention to purchase'? Not so, said Marx, who added that a capitalist may sell in order to pay, especially during a crisis. And:

> During the crisis, a man may be very pleased if he has sold his commodities without immediately thinking of a purchase … The immediate purpose of capitalist production is not 'possession of other goods' but the appropriation of value, of money, of abstract wealth. (Marx 1981, p. 503)

In the circulation of capital, M–C–M, the possibility of crisis is developed to its fullest extent. Firstly, it is a development of the 'simple' circulation of commodities, C–M–C, and therefore contains the 'simple' possibility of crisis, namely that commodities must (yet may not be able to) undergo a sequence of transformations. Secondly, under capitalist production, money as means of payment introduces a far-reaching set of connections between capitals. Thirdly, the fact that the goal of capitalist production is the acquisition of abstract wealth, rather than other use-values, means that the presence of use-values for sale is no longer sufficient to ensure that sales will take place, let alone at prices which will give the desired return.

Marx's criticism of Ricardo has a wider significance. Ricardo was criticized here not for erring in his deductions, but rather because the starting point for those deductions, his 'model', was inappropriate. Leaving aside those unfortunate moments when he was using arguments relevant only to a barter economy, Ricardo's model was one of simple commoditiy production, characterized by the circuit C–M–C. This was inappropriate, said Marx, because the circulation of capital, M–C–M, contains new possibilities for crises, not contained in the simple circulation C–M–C.

If Marx was right about this, then any model of production and exchange where the objective is consumption (that is, the acquisition of use-values rather than value in general) by its very nature excludes those specifically *capitalist* causes of the possibility of crisis.

The converse of this is that a proper consideration of capitalist crisis must consider not only use-values but value too: 'value, abstract wealth, money'. In this respect, Keynes's introduction of effective demand into the orthodox theory of his time can be seen as an attempt to remedy the same

one-sidedness of that theory which Marx criticized in Ricardo. Indeed, the theory of the possibility of crisis can help show why 'effective demand' – a monetary quantity – is important in its own right and why Keynes was justified in elevating it to a place of considerable importance (Kenway 1980).

Ricardo denied that crises could arise out of the production process itself. In his defence, Marx commented that Ricardo himself did not actually experience any such crises (Marx 1861, p. 497). All the crises between 1800 and 1815 could be attributed to external conditions: poor harvest; interference with the currency by the authorities; the wars. After 1815, the crises could be explained quite readily by reference to the strains of the change from war to peace. Yet as Marx observed, these interpretations were not available to Ricardo's followers. And neither, of course, are they available today.

See Also

- ▶ Business Cycles
- ▶ Marxist Economics
- ▶ Trade Cycle

References

Kenway, P.M. 1980. Marx, Keynes and the possibility of crisis. *Cambridge Journal of Economics* 4(1): 23–36.
Marx, K. 1861. *Theories of surplus value*, Part 2. London: Lawrence & Wishart, 1969.
Marx, K. 1867. *Capital*, vol. I. Harmondsworth: Penguin, 1976.
Ricardo, D. 1821. In *Collected works and correspondence*, vol. I, ed. P. Sraffa. Cambridge: Cambridge University Press, 1951.

Critical Path Analysis

Kenneth R. MacCrimmon

When consumers plan vacations, manufacturers schedule production, and governments tackle budget deficits, each must deal with a myriad of interrelated activities. In large projects, managing these interrelationships is very difficult due to three major factors: the precedence ordering of activities, the uncertainty about activity durations, and the possibility of reallocating resources.

Even if there were only one way to perform an activity and if the time it took to complete it were known for sure, there would still be the problem of determining when an activity can begin. A book cannot be bound until it is printed, cannot be printed until it is edited, and cannot be edited until it is written. When one realizes that there are hundreds of activities in book publishing, it is clear that effective management requires some way of keeping track of the precedence order of activities.

A further complexity is the uncertain duration of activities. People get sick, buildings burn down, funds are scarce and so activities often take longer than expected. Although delays in some activities will be relatively unimportant, delays in others will delay the whole project. Effective project management must focus attention on such critical activities.

The third major complication is due to the multiplicity of ways in which things can be done. By allocating more resources, an activity can be speeded up. By allocating fewer resources, costs can be held down, although delays will probably occur. Project managers need a way of determining how resources can be effectively allocated.

Critical path analysis (CPA) is the generic name for a set of techniques to help people deal with the problems of managing projects. The basic elements in CPA are: (i) the specification of activities necessary to complete a project; (ii) their precedence order represented by a directed, acyclic network diagram; (iii) the identification of the critical activities, especially those activities on the longest path through the network (i.e., the critical path); and (iv) the determination of cost–time tradeoffs for the whole project.

Critical path analysis can be viewed as the consolidation and extension of the ideas of Henry Gantt and Vilfredo Pareto. In the early 1900s, Gantt proposed a graphical method for scheduling and controlling production activities.

A Gantt chart represents each production activity as a row with a bar drawn to scale representing how long the activity takes. The bar is positioned in calendar time by taking into account the precedence relationships among activities, although these relationships are not shown directly.

Pareto suggested that a small proportion of components had an undue effect on the performance of a system. For example, 10 per cent of a company's sales force often account for 90 per cent of the sales. This concept, sometimes under the label of the 'Pareto principle', became adopted as a regular management control technique. Both John Commons and Chester Barnard incorporated this concept of the 'critical factor' or 'strategic factor' in their economic theories of organization.

In the late 1960s two independent techniques, PERT (Program Evaluation and Review Technique) and CPM (Critical Path Method), were created to help manage very large projects. PERT was used in handling the tens of thousands of activities in developing weapon systems for Polaris submarines for the US Navy (Malcolm et al. 1959). CPM was developed for controlling large construction projects in industry (Kelley and Walker 1959). Both methods can be viewed as a Gantt chart embedded in a network that shows the interdependencies among activities. This representation shows how the expected start and completion times of any activity depend on the progress of the activities that precede it. By identifying which activities can delay the whole project and which can expedite the project, the concept of critical factors is developed into a concept of a 'critical path'. By associating time–cost trade-offs with each activity, it becomes clear that one speeds up a project by allocating resources to critical activities and one saves money by withdrawing resources from non-critical activities.

Even though both PERT and CPM used very similar ideas in network representation and the identification of critical paths, each technique had its own unique features. In PERT, the arcs in a network represented the activities and the nodes represented starting and ending points, while in CPM the nodes represented the activities and the arcs indicated the precedence relationships. A more significant difference was that PERT allowed for uncertainty in the duration of an activity while CPM exhibited time-resource trade-offs. Since both uncertainty and resource trade-offs are key elements of any large project, both PERT and CPM made distinctive contributions.

PERT assumed that the uncertainty in the duration of an activity could be represented by a Beta distribution, the parameters of which are derived from three time estimates, a most likely time, an optimistic time, and a pessimistic time provided by project managers. Using the Central Limit Theorem, the overall project time is normally distributed with a mean (variance) equal to the sum of the means (variances) of activity distributions along the critical path. The possible errors in the PERT assumptions, at both the level of individual activities and at the level of the whole project have been analysed (MacCrimmon and Ryavec 1959). The assumption of independence among activities (allowing means and variances to be summed) is particularly weak. Environmental events that delay one activity are likely to delay other activities and so the estimated completion time will tend to be optimistic. Methods have been proposed for grouping network elements to reduce bias and for using simulation techniques to overcome some of the analytical difficulties.

While PERT focuses on time management, CPM focuses on the cost of performing activities. Piecewise linear time–cost tradeoffs are developed from information provided by project managers. For any desired project completion time, linear programming can be used to ascertain the minimum project variable cost subject to resource availabilities. By varying the project completion time parameter, a frontier of tradeoffs of total variable cost and completion times is obtained. By focusing only on the cost of the resources, the allocation of specific resources remains to be determined separately. The resources for speeding up one activity may be committed elsewhere.

Clearly, both PERT and CPM can help to plan and control large projects. By combining the best features of each, there is promise for developing more powerful methods. One of the first modifications was the development of PERT–COST which took into account project costs, although

in more of a monitoring role than in CPM. Over time the original distinctions between PERT and CPM have become blurred and it is reasonable to focus on a generic CPA. Advances have taken place in two main areas, handling uncertainty and managing resources.

Handling project uncertainty has been improved by de-emphasizing the single most critical path. When delays occur, other paths may become the critical path, thus activities that are common to several of these paths should be monitored carefully. Methods allowing for a more flexible treatment of uncertainty in the activity durations have also been developed (Elmaghraby 1977).

There is also uncertainty in how a project can be carried out. As the project goes along, the results of early activities influence the way later activities are performed. For example, the outcome of research and development on a new kind of memory may have major implications for the construction of a computer. More advanced network models, then, allow for uncertainty in network structure such as disjunctive activities whereby one activity is performed in lieu of another (Pritsker and Sigal 1983).

A second major area of improvement has been in handling resources (Dean and Chaudhuri 1980). Procedures for resource smoothing were used to avoid costly fluctuations such as continual hiring and firing in the labour force. Methods have been proposed for splitting jobs, allowing for halting the performance of one activity and transferring resources to where they are most needed. More sophisticated techniques have been developed for handling multiple categories of resource types and for handling uncertainty about the availability of resources. Other advances have incorporated information about the quality of performance of the activity.

Critical path analysis is now widely used in one form or another. Many actual applications, however, involve only the most basic elements such as the network representation. Why aren't some of the more advanced methods used? As one manager is reported to have said about the time estimates required in PERT, 'activity durations are too uncertain to try to use more than one time estimate'! With better analytical training, with microcomputers, and with sophisticated computer programs, perhaps the uses of critical path analysis will begin to catch up with the developments in the methods.

See Also

- ▶ Combinatorics
- ▶ Operations Research

Bibliography

Dean, B.V., and A.K. Chaudhuri. 1980. Project scheduling: A critical review. *TIMS Studies in the Management Sciences* 15: 215–233.

Elmaghraby, S.E. 1977. *Activity networks: Project planning and control by network models*. New York: Wiley.

Kelley, J.E., and M.R. Walker. 1959. Critical path planning and scheduling. *Proceedings of the Eastern Joint Computer Conference*.

MacCrimmon, K.R., and C.A. Ryavec. 1959. An analytical study of the PERT assumptions. *Operations Research* 7(5): 16–37.

Malcolm, D.G., J.H. Rosenboom, C.E. Clark, and W. Fazar. 1959. Applications of a technique for research and development program evaluation. *Operations Research* 7(5): 646–669.

Pritsker, A.A.B.., and C.E. Sigal. 1983. *Management decision making: A network simulation approach*. Englewood Cliffs: Prentice-Hall.

Croce, Bendetto (1866–1952)

R. Bellamy

Croce was a southern Italian idealist philosopher and historian. His *Philosophy of Spirit* was intended as a secular religion capable of encompassing all aspects of human life. He regarded as his greatest innovation the addition of the category of the Useful to the classical triad of the Beautiful, the True and the Good. He elaborated this theory in the course of his early

writings on Marx (1900b) and a debate with Pareto 'On the Economic Principle' (1900a). He argued that human practical activity was orientated to solving the immediate problems of everyday life, and hence highly contingent. We only discover the moral worth of an act post facto, when the consequences can be evaluated. Our action is therefore directed at the Useful and only indirectly at the Good. Whilst all moral acts are economic, the reverse is not the case. He rejects hedonism and egoism as ethical theories, since happiness and self-interest may be good guides to the utility of an act to an agent at a given time, but not necessarily to its ultimate moral worth. He therefore disputed Pareto's contention that you could develop a science of economics based on certain constant features of human behaviour. All human activity is conditioned by chance and the diversity of beliefs different individuals hold. This fact similarly vitiated Marx's historical materialism. These ideas were later expanded into his *Philosophy of the Practical* (1908). However, in a later debate he denied Luigi Einaudi's conclusion that his theory implied classical liberal laissez-faire policies (1928). He asserted that certain conditions could warrant welfare socialism. A moderate conservative rather than a liberal, he belatedly opposed fascism, partly because his philosophy provided few action guiding principles in the present beyond the endorsement of whatever succeeds. The judgement of events is left to history.

See Also

► Pareto, Vilfredo (1848–1923)

Selected Works

1900a. On the economic principle: A letter to Professor V. Pareto. In *International Economic Papers*, ed. A.T. Peacock et al. London: Macmillan, 1953.

1900b. *Historical materialism and the economics of Karl Marx*. Trans. C.M. Meredith, London: Howard Latimer Ltd, 1931.

1908. *Philosophy of the practical*. Trans. D. Ainslee. London: Macmillan, 1913.

1928. Free enterprise and liberalism. In B. Croce, *Politics and morals*. Trans. S.J. Castiglione. London: George Allen & Unwin, 1946.

References

Bellamy, R.P. 1986. *Modern Italian social theory*. Cambridge: Polity Press. ch. 5.

Crosland, Anthony (1918–1977)

I. M. D. Little

Born in 1918, Crosland read classics at Trinity College, Oxford (1937–40). Always a socialist, he led the undergraduate faction that opposed the Communist creed embraced by many left-wing intellectuals of that period. War service, as a paratrooper, claimed him for the next five years. Returning to Oxford, he became President of the Union, took a first class degree in politics and economics, and was appointed to a Fellowship at Trinity College. He taught economics for three years, then in 1950 began a political career as a Labour Member of Parliament.

His most important book, *The Future of Socialism* (1956), sought to define the role of a Socialist government in a modern industrial state. Essentially anti-utopian and revisionist, it was as opposed to latter-day Marxism as to Toryism. Crosland insisted that Socialism was about equality, not the ownership of the means of production. Greater equality was facilitated by economic growth, and high levels of government expenditure and intervention were also required.

For 20 years his views strongly influenced the Labour Party. Between 1964 and his sudden death in 1977 when he had been Foreign Secretary only 10 months, he held four senior Cabinet posts in which he initiated measures born of his political philosophy, particularly in education and housing.

But government failure to achieve sufficient growth frustrated many of his aspirations. And after his death his blueprint for the future of socialism became less realizable. Always an optimist, he underestimated the economic and social forces that would obstruct his programmes.

See Also

- ▶ Fabian economics
- ▶ Social Democracy

Selected Works

1956. *The Future of Socialism.* London: Jonathan Cape.

Cross-cultural Experiments

Rob Boyd

Abstract

Experiments conducted in student populations suggest that people are not money maximizers, but also seem to have social preferences. To determine whether these social preferences are culturally variable, a group of economists and anthropologists undertook a series of economic experiments in a wide range of non-Western, small scale societies. Results in these societies were highly variable, and in some of them strikingly different from experiments in student populations. Variation in behaviour was correlated with societal characteristics, but not individual attributes. Finally, variation in punishment across societies predicted variation in cooperation across societies.

Keywords

Altruism; Cooperation; Culture and economics; Dictator game; Economic experiments; Experimental games; Group characteristics; Public goods games; Reciprocity; Social preferences; Third-party punishment game; Ultimatum game

JEL Classifications

C9

A large number of well-replicated results using a wide variety of experimental games are inconsistent with the assumption that people are money maximizers. Instead, people's behaviour is consistent with choices based on social preferences in which people place a positive value on fairness, reciprocity, or equity (see Camerer 2003, for a review). For example, subjects typically make significant positive contributions in the public goods games, reject positive offers in the ultimatum game, and impose costly punishment in the third-party punishment game (see Camerer 2003, ch. 2, for descriptions of these games.) In some games these results are insensitive to framing and whether behaviour is anonymous to the experimenter ('double blind').

These experiments are open to two qualitatively different interpretations: It could be that pro-social behaviours like cooperation in the public goods game and punishment in the third-party punishment game reflect human nature. Cooperation in the public goods game could result from universal cognitive systems that cause people everywhere to behave as if all acts have reputational consequences, even when facts suggest no one will know what they have done. Punishment in the third-party punishment game could result from a pan-human motivational system that causes people to prefer outcomes that are fair or mutually beneficial, and to derive satisfaction from punishing unfair behaviour. However, with few exceptions experimental subjects have been university students in urbanized, industrial societies. Thus, it also could be that observed pro-social behaviour results from culturally evolved beliefs and values that are specific to such social environments. It is obviously of great importance to determine which of these two interpretations is correct.

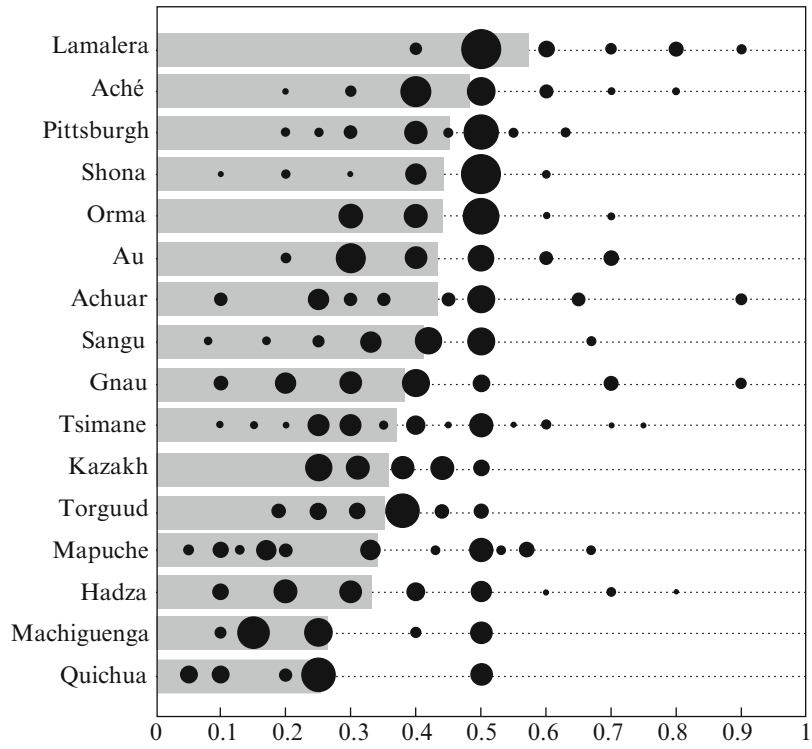
To answer this question, a team of anthropologists and economists performed two rounds of

experimental games in a wide range of cultural environments. The first round (Henrich et al. 2004, 2005) comprised a diverse group of 15 societies including peoples like the Aché and Hadza who live in nomadic foraging bands, the Achuar and Au who live in small villages and mix hunting and horticulture, Mongol and Sangu pastoralists, and sedentary Shona farmers in Zimbabwe. The ultimatum game was performed in all 15 societies, and the public goods game and the dictator game were performed in different subsets. The second round (Henrich et al. 2006) included a similar and overlapping range of 15 societies. Based on experience in the first round, experimental protocols were improved and standardized, and a greater effort was made to collect standardized data on individual characteristics. During the second round the ultimatum, dictator, and third-party punishment games were performed in all 15 societies. In addition complete strategies for second players in the ultimatum game and punishers in the third-party punishment game were elicited using the strategy method.

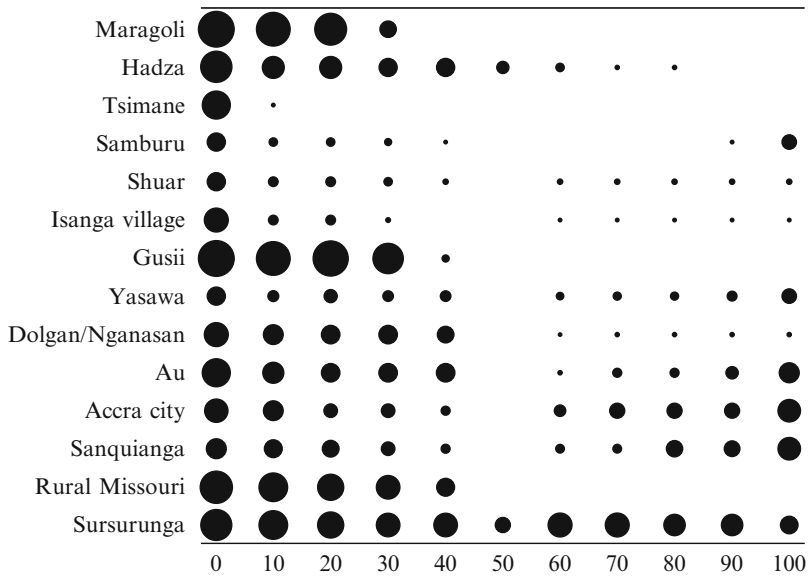
These experiments reveal a number of interesting results.

1. *Behaviour in non-Western populations can be quite different from that of Western university subjects.* Figure 1 shows the distribution of ultimatum game offers in the first round of experiments. The Pittsburgh data taken from Roth et al. (1991) are typical for university populations – the modal offer is 50 per cent but many subjects make somewhat lower offers. Behaviour in other populations can be very different. For example, modal offers are much lower among two lowland tropical forest groups; the Achuar and the Machiguenga are quite low. Interestingly, these very low offers were usually accepted, behaviour much closer to the predictions of money maximization than the behaviour of Western university subjects. Non-western populations also exhibited novel behaviours not seen in university populations.

   Figure 2 shows the rejection probabilities for different ultimatum game offers. Notice



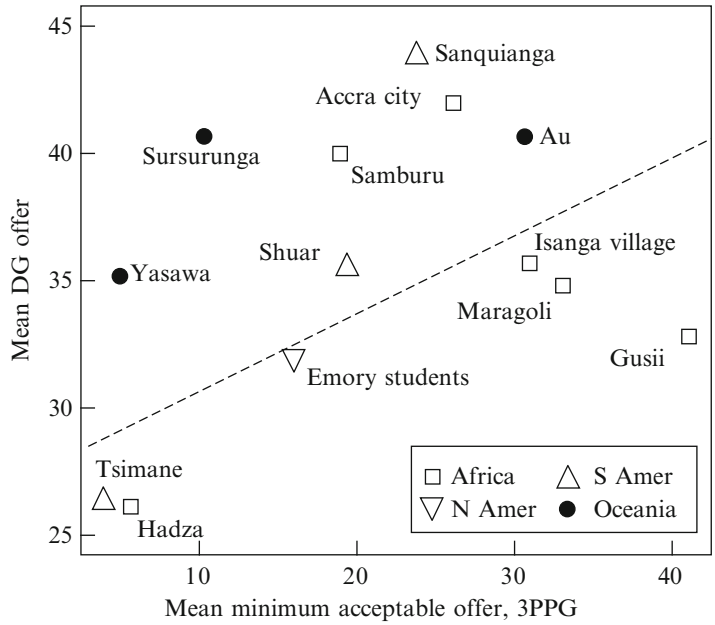
Cross-cultural Experiments, Fig. 1 Ultimatum game offer. *Note:* A *bubble plot* showing the distribution of ultimatum game offers for each group. The diameter of the circle at each location along each row represents the proportion of the sample that made a particular offer. The *right edge* of the *lightly shaded horizontal grey bar* is the mean offer for that group. In the Machiguenga row, for example, the mode is 0.15, the secondary mode is 0.25, and the mean is 0.26. *Source:* Henrich et al. (2005)



Cross-cultural Experiments, Fig. 2 Ultimatum game rejection rates. *Note:* The diameter of the *black circles* is proportional to the fraction of offers that would have been rejected in the ultimatum game during the second round of experiments plotted as a function of the offer as a percentage of the maximum offer. For scale, note that the Gusii and Maragoli rejected all offers of zero. Notice that in all societies offering 50% of the stake minimizes the probability of rejection, but that in a number of societies increasing offers above 50% increases the rate of rejection. *Source:* Henrich et al. (2006)

that in several populations increasing offer level above 50 per cent *increased* the rate of rejections, a phenomenon not observed among student subjects.

2. *Behavioural differences are correlated with group characteristics but not individual characteristics.* The ethnographers who performed most of these experiments have studied these groups for many years and have detailed data on subjects about income, wealth, education, market contact, and a variety of other factors. None of these factors was significantly correlated with ultimatum game offers within social groups in first round, or offers or rejections in the second round. Because measures of wealth, income, and so on are not comparable across groups, these measures could not be aggregated to derive group characteristics. However, during the first round, ethnographers who were blind to the results ranked each of the groups along five dimensions: extent of cooperation in subsistence, degree of market contact, amount of privacy, amount of anonymity, and social complexity. We also had comparable data on settlement size. It turned out that market contact, settlement size, and social complexity were all highly correlated, so these were collapsed into a single variable labelled 'aggregate market contact'. Multiple linear regression showed that increasing aggregate market contact and cooperation in subsistence significantly predicted increased ultimatum game offers, and together the two variables accounted for more than half of the variance among groups in average offers.
3. *Variation in punishment predicts variation in altruism across societies.* In the third-party punishment game, an individual, the 'punisher' observes a dictator game and can punish the dictator at a cost to him or herself. The average minimum offer acceptable to the punisher in this game provides a measure of the level of punishment in that society. As is shown in Fig. 3, this measure of punishment also predicts the level of altruism measured by dictator offers in the ordinary dictator game.



Cross-cultural Experiments, Fig. 3 Mean minimum acceptable offer, third-party punishment game. *Note:* The mean offer in the dictator game for a society plotted against the mean value of the minimum acceptable offer in the third-party punishment game. The different symbols indicate continents. The size of each symbol is proportional to the number of DG pairs at each site. The *dotted line* gives the weighted regression line, with continental controls of mean dictator game offers against mean minimum acceptable offer in the third-party punishment game. *Source:* Henrich et al. (2006)

Taken together these results indicate that pro-social behaviour in economic experiments does not result from an invariant property of our species, and instead suggest that there are significant cultural differences between societies. The fact that ultimatum game behaviour is predicted by the average level of cooperation and average level of market contact further indicates that these cultural differences are not arbitrary, but may reflect economic, ecological and social differences between societies. However, the lack of correlation between individual characteristics and individual behaviour indicates that the differences between societies are not likely to be explained as the simple aggregation of individual experiences. Instead, it is more plausible that cultures evolve over time in response to the average conditions which they face, and that individual behaviour is, in turn, shaped by these cultural differences.

See Also

► Experimental Economics

Bibliography

Camerer, C. 2003. *Behavioral game theory: Experiments on strategic interaction*. Princeton: Princeton University Press.

Henrich, J.R.B., S. Bowles, C. Camerer, E. Fehr, and H. Gintis. 2004. *The foundations of human sociality: Economic experiments and ethnographic evidence from fifteen small-scale societies*. New York: Oxford University Press.

Henrich, J., R. Boyd, S. Bowles, C. Camerer, E. Fehr, H. Gintis, R. McElreath, M. Alvard, A. Barr, J. Ensminger, K. Hill, F. Gil-White, M. Gurven, F. Marlowe, J.Q. Patton, N. Smith, and D. Tracer. 2005. 'Economic man' in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences* 28: 795–855.

Henrich, J., R. McElreath, A. Barr, J. Ensminger, C. Barrett, A. Bolyanatz, J.C. Cardenas, M. Gurven, E. Gwako, N. Henrich, C. Lesorogol, F. Marlowe,

D. Tracer, and J. Ziker. 2006. Costly punishment across human societies. *Science* 312: 1767–1770.
Roth, A.E., V. Prasnikar, M. Okuno-Fujiwara, and S. Zamir. 1991. Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An experimental study. *American Economic Review* 81: 1068–1095.

C

Crowding Out

Olivier Jean Blanchard

Abstract

'Crowding out' refers to all the things which can go wrong when debt-financed fiscal policy is used to affect output. While the initial focus was on the slope of the LM curve, 'crowding out' now refers to a multiplicity of channels through which expansionary fiscal policy may in the end have little, no or even negative effects on output.

Keywords

Accumulation of capital; Aggregate demand; Budget deficits; Crowding out; Fiscal consolidation; Fiscal expansion; Fiscal policy; Flexible exchange rates; Full employment; Inflation; Interest rates; Intertemporal taxation; Investment tax credit; IS–LM model; Labour supply; Lump sum taxes; Multiplier analysis; Mundell–Fleming model; Natural rate of unemployment; Private spending; Public debt; Public expenditure; Ricardian equivalence theorem; Risk premium; Taxation of income

JEL Classifications

E2

'Crowding out' refers to all the things which can go wrong when debt-financed fiscal policy is used to affect output.

A first line of argument questions whether fiscal policy has any effect at all on spending. Changes in the pattern of taxation which keep the pattern of spending unaffected do not affect the intertemporal budget constraint of the private economy and thus may have little effect on private spending. This argument, known as the 'Ricardian equivalence' of debt and taxation, holds only if taxes are lump sum (Barro 1974). Some taxes which induce strong intertemporal substitution, such as an investment tax credit for firms, will have stronger effects if they are temporary; for most others, such as income taxes, changes in the intertemporal pattern may have only a small effect on the pattern of spending.

The Ricardian equivalence argument is not settled empirically and its validity surely depends on the circumstances. A change in the intertemporal taxation of assets such as land or housing, leaving the present value of taxes the same, will have little effect on their market value, thus on private spending. An explicitly temporary income tax increase may have little effect on spending while the anticipation of prolonged deficits may lead taxpayers to ignore the eventual increase in tax liabilities. Evidence from specific episodes, such as the 1968 temporary tax surcharge in the United States, suggests partial offset at best.

Changes in the pattern of government spending obviously have real effects. But here again, various forms of direct crowding out may be at work. Public spending may substitute perfectly or imperfectly for private spending, so that changes in public spending may be directly offset, fully or partially, by consumers or firms. Even if public spending is on public goods, the effect will depend on whether the change in spending is thought to be permanent or transitory. Permanent changes, financed by a permanent increase in taxes, will, as a first approximation, lead to a proportional decrease in private spending, with no effect on total spending. Temporary changes in spending, associated with a temporary increase in taxes, lead to a smaller reduction in private spending and thus to an increase in total spending.

In summary, one should not expect any change in taxation or government spending to have a one-for-one effect on aggregate demand. An eclectic reading of the discussion above may be that only

sustained decreases in income taxation, or the use of taxes that induce strong intertemporal substitution, or temporary increases in spending, can reliably be used to boost aggregate demand. The focus in what follows will be on these forms of fiscal expansion.

Crowding Out at Full Employment

Not every increase in aggregate demand translates into an increase in output.

This is clearly the case if the economy is already at full employment (I use 'full employment' to mean employment when unemployment is equal to its natural rate). While tracing the effects of fiscal expansion at full employment is of limited empirical interest, except perhaps as a description of war efforts, it is useful for what follows. If labour supply is inelastic, output is fixed and any increase in aggregate demand must be offset by an increase in interest rates, leaving output unchanged. In the case of an increase in public spending, private spending will decrease; in the case of a decrease in income taxation, private spending will in the end be the same, but its composition will change as the share of interest sensitive components decreases. (If labour supply can vary, the story is more complicated. See, for example, Baxter and King 1993, for an analysis of changes in government spending in an otherwise standard RBC model.)

This is just the beginning of the story, however. Over time, changes in capital and debt lead to further effects on output. The decrease in investment in response to higher interest rates leads to a decline in capital accumulation and output, reducing the supply of goods. If fiscal expansion is associated with sustained deficits, the increase in debt further increases private wealth and private spending at given interest rates, further increasing interest rates and accelerating the decline in capital accumulation (see, for example, Blanchard 1985, for a characterization of these dynamic effects in an economy with finite horizon consumers). How strong is this negative effect of debt on capital accumulation likely to be? One of the crucial links in this mechanism is the effect of government debt on interest rates; empirical evidence, both across countries and from the last two centuries, shows surprisingly little relation between the two. This probably reflects, however, more the difficulty of identifying and controlling for other factors than the absence of an effect of debt and deficits on interest rates.

Worse can happen. It may be that the fiscal programme becomes unsustainable. There is no reason to worry about a fiscal programme in which debt grows temporarily faster than the interest rate. But there is reason to worry when there is a positive probability that, even under the most optimistic assumptions, debt will have to grow for ever faster than the interest rate. When this is the case, it implies that the government can meet its interest payments on existing debt only by borrowing more and more. What happens then may depend on the circumstances. Bond holders may start anticipating repudiation of government debt and require a risk premium on the debt, further accelerating deficits and the growth of the debt. If they instead anticipate repudiation through inflation, they will require a higher nominal rate and compensation for inflation risk in the form of a premium on all nominal debt, private and public. What is sure is that there will be increased uncertainty in financial markets and that this will further contribute to decreases in output and in welfare. The historical record suggests that it takes very large deficits and debt levels before the market perceives them as potentially unsustainable. England was able in the 19th century to build debt-to-GDP ratios close to 200 per cent without apparent trouble. Some European countries are currently running high deficits while already having debt-to-GDP ratios in excess of 100 per cent, without any evidence of a risk premium on government debt. The threshold seems lower for Latin American economies. But even if one excludes this worst-case scenario, fiscal expansion can clearly have adverse effects on output at full employment. The relevant issue, however, is whether the same dangers are present when fiscal expansion is implemented to reduce unemployment, which is presumably when it is most likely to be used.

Crowding Out at Less Than Full Employment

The historical starting point of the crowding out discussion is the fixed price IS–LM model. In that model, a fiscal expansion raises aggregate demand and output. The pressure on interest rates does not come from the full employment constraint as before but from the increased demand for money from increased output. Thus the fiscal multiplier is smaller the lower the elasticity of money demand to interest rates, or the larger the elasticity of private spending to interest rates. Fiscal expansion crowds out the interest-sensitive components of private spending, but the multiplier effect on output is positive. As output and interest rates increase, it is quite possible for both investment and consumption to increase. But what happens when the model is extended to take into account dynamics, expectations and so on? Can one overturn the initial result and get full crowding out or even negative multipliers?

Even within the static IS–LM, one can in fact get zero or negative multipliers. This is the case, for example, if money demand from agents is higher than that from the government and the change in policy redistributes income from the government to agents. While this case is rather exotic, a much stronger case can be made if the economy is small, open, and with capital mobility and flexible exchange rates, as in the 'Mundell–Fleming' model. In this case, with the interest rate given from outside, and fixed money supply, money demand determines output; fiscal policy leads only to exchange rate appreciation. Exchange rate-sensitive components are now crowded out by fiscal expansion. The multiplier is equal to zero.

When dynamic effects are taken into account, other channels arise for crowding out. The analysis of these dynamic effects, with the dynamics of debt accumulation taken into account, was initially conducted under the maintained assumption of fixed prices and demand determination of output (Tobin and Buiter 1976). Then, as debt was accumulating, private wealth and spending increased, leading to even larger effects of fiscal policy on output in the long run than in the short run. But the assumption of fixed prices, while debt and capital accumulation are allowed to proceed, is surely misleading; when prices are also allowed to adjust, the effects of fiscal policy become more complex, and crowding out more likely. This is because some of the full employment effects come back into prominence: if fiscal expansion is maintained even after the economy has reached full employment, then the perverse effects of higher interest rates on capital accumulation and full employment output come again into play. This is true even if deficits disappear before the economy returns to full employment; the economy inherits a larger level of debt, and thus must have higher interest rates and lower capital accumulation than it would otherwise have had. The fiscal expansion trades off a faster return to full employment for lower full-employment output.

Anticipations of these full employment effects are likely to feed back and modify the effects of fiscal policy at the start, when the economy is still at less than full employment. Anticipations of higher interest rates, perhaps also of higher distortions due to the higher taxes needed to service the debt, may dominate the direct effects of higher government spending on demand, and lead to an initial decrease rather than an initial increase in demand and output. Symmetrically, fiscal consolidation, to the extent that it implies lower interest rates and lower distortions in the future, may be expansionary. This is even more likely to be the case if fiscal consolidation decreases the risk of default on government debt, and thus decreases the risk of major economic disruptions. There is indeed some evidence that, when initial fiscal conditions are very bad, and the fiscal consolidation is large and credible, the net effect of consolidation may be expansionary (Giavazzi and Pagano 1990).

Crowding Out: An Assessment

Should one conclude from this that fiscal policy is an unreliable macroeconomic tool, with small and sometimes negative effects on output? The answer is 'no'. Fiscal policy is likely to partly crowd out some components of private spending, even in the

best circumstances, but there is little reason to doubt that it can help the economy return to full employment. Ricardian equivalence and direct crowding out warn us that not any tax cut or spending increase will increase aggregate demand. But there is little question that temporary spending or sustained income tax cuts will do so. Results of full crowding out at less than full employment, such as the Mundell-Fleming result, are simply a reminder that the monetary-fiscal policy mix is important.

In all cases, monetary accommodation of the increased demand for money removes the negative or the zero multipliers. That fiscal expansion affects capital accumulation, and output adversely at full employment, and that unsustainable fiscal programmes may lead to crises of confidence, is a reminder that fiscal expansion should not be synonymous with steady increases in the debt-to-GDP ratio even after the economy has returned to full employment. This shows one of the difficulties associated with fiscal expansion: if done through tax cuts, it has to be expected to last long enough to affect private spending, but not so long as to lead to expectations of runaway deficits in the long run. The room for manoeuvre is, however, substantial. Some taxes, such as the investment tax credit, work best when temporary. These can be used, as they work in the short run and have few adverse implications for the long run.

See Also

- ▶ Budget Deficits
- ▶ Real Business Cycles
- ▶ Ricardian Equivalence Theorem
- ▶ Tobin, James (1918–2002)

Bibliography

Barro, R. 1974. Are government bonds net wealth? *Journal of Political Economy* 82: 1095–1117.

Baxter, M., and R. King. 1993. Fiscal policy in general equilibrium. *American Economic Review* 83: 315–334.

Blanchard, O. 1985. Debt, deficits, and finite horizons. *Journal of Political Economy* 93: 223–247.

Giavazzi, F., and M. Pagano. 1990. Can severe fiscal contractions be expansionary? *NBER Macroeconomics Annual* 5: 75–122.

Tobin, J., and W. Buiter. 1976. Long-run effects of fiscal and monetary policy on aggregate demand. In *Monetarism*, ed. J. Stein. Amsterdam: North-Holland.

Crowther, Geoffrey (1907–1972)

R. J. Bigg

Crowther was educated at Leeds Grammar School, Oundle and Clare College, Cambridge, where after studying modern languages he proceeded to win a high first in Part II of the Economics Tripos. Lionel Robbins remembered one of his examination answers as being only a few sentences: 'the way he put it left nothing more to be said' (Robbins 1972, p. 23). This ability to go to the heart of the matter Crowther carried into his work at *The Economist*.

Prior to this however he went to Yale and Columbia Universities as a Commonwealth Fund Fellow (he married an American, Margaret Worth, in 1932) and then worked for two years in a London merchant bank. This led to his appointment as economic adviser on banking to the Irish government. Crowther gave up the Irish appointment to join *The Economist* in 1932, becoming assistant editor in 1935 and editor in 1938. Crowther was the longest major editor of the newspaper, holding the post from 1938 to 1956 – Robbins compared him to the paper's previous great editor, Walter Bagehot. After 1956 he maintained his contact with the journal, first as Managing Director and later as Chairman.

Under Crowther *The Economist* changed radically in its format so as to widen its appeal to a broader readership both in the UK and overseas, expanding its circulation from 10,000 to 55,000 and becoming one of the most influential weekly papers in the world. Crowther was also responsible for the establishment of the Economist Intelligence Unit just after World War II and for the

newspaper's successful development of its St James's property. He was also one of the first newspaper editors to appoint women, such as Barbara Ward, to the staff in significant positions.

His major theoretical work on economics was *An Outline of Money* (1940), which, like his journalistic writings, had 'a clarity and expository power which few academics could muster' (Robbins 1972, p. 23) as well as a lively style and was quickly popular with both students and the general public. His other works on economics, such as *Ways and Means* (1936) stemmed from broadcasts or lectures.

Crowther's magnetic personality and prodigious capacity for work made him an outstanding public servant (Goode 1974). His principal public interest was in education. He was Chairman of the Central Advisory Council for Education (England) from 1956 to 1960, whose report '15 to 18' was a landmark in the expansion of further education. As the first Chancellor of the Open University (1968) he played a major part in its early development. He then took on the joint responsibilities of chairing both the Royal Commission on the Constitution and the Committee on Consumer Credit, whose report (1971) recommended the complete reform of consumer credit law (embodied in the Consumer Credit Act of 1974) and of personal property security law. Crowther was knighted in 1957 for his services to journalism and in 1968 he became a life peer, taking his title from Headingley, the place of his birth.

Selected Works

1936. *Ways and means*: *A study of the economic structure of Great Britain today*. London: Macmillan.
1940. *An outline of money*. London: Thomas Nelson & Sons.

References

Goode, R.M. 1974. A credit law for Europe? *International and Comparative Law Quarterly* 23(1): 227–291.
Robbins, L.S. 1972. Lord Crowther: Memorial address. *The Economist*. 25 Mar.

Cryptocurrency

Eli Dourado and Jerry Brito

Abstract

For most of history, humans have used commodity currency. Fiat currency is a more recent development, first used around 1000 years ago, and today it is the dominant form of money. But this may not be the end of monetary history. Cryptocurrency is neither commodity money nor fiat money – it is a new, experimental kind of money. The cryptocurrency experiment may or may not ultimately succeed, but it offers a new mix of technical and monetary characteristics that raise different economic questions than other kinds of currency.

This article explains what cryptocurrency is and begins to answer the new questions that it raises. To understand why cryptocurrency has the characteristics it has, it is important to understand the problem that is being solved. For this reason, we start with the problems that have plagued digital cash in the past and the technical advance that makes cryptocurrency possible. Once this foundation is laid, we discuss the unique economic questions that the solution raises.

Keywords

Anonymity; Bitcoin; Byzantine Generals Problem; Censorship resistance; Cryptocurrency; Cryptography; Double spending problem; Exchange rate indeterminacy; Mining pools; Money; New monetary economics; Open source; Peer-to-peer networking; Proof of work; Pseudonymity; Trust; Volatility

JEL Classifications

E40; L31; F31; O30

Technical Overview

Cryptocurrency is the name given to a system that uses cryptography to allow the secure transfer and exchange of digital tokens in a distributed and decentralised manner. These tokens can be traded at market rates for fiat currencies. The first cryptocurrency was Bitcoin, which began trading in January 2009. Since then, many other cryptocurrencies have been created employing the same innovations that Bitcoin introduced, but changing some of the specific parameters of their governing algorithms. The two major innovations that Bitcoin introduced, and which made cryptocurrencies possible, were solutions to two long-standing problems in computer science: the double-spending problem and the Byzantine Generals Problem.

Double Spending

Until the invention of Bitcoin, it was impossible for two parties to transact electronically without employing a trusted third party intermediary. The reason was a conundrum known to computer scientists as the 'double spending problem', which has plagued attempts to create electronic cash since the dawn of the Internet.

To understand the problem, first consider how physical cash transactions work. The bearer of a physical currency note can hand it over to another person, who can then verify that he is the sole possessor of that note by simply looking at his hands. For example, if Alice hands Bob a $100 bill, Bob now has it and Alice does not. Bob can easily verify his possession of the $100 bill and, implicitly, that Alice no longer has it. Physical cash transfers are also final, in the sense that to reverse a transaction the new bearer must give back the currency note. In our example, Bob would have to hand the $100 bill back to Alice. Given all of these properties, cash makes it possible for different parties, including strangers, to transact without trusting each other.

Now, consider how electronic cash might work. Obviously, paper notes would be out of the picture. There would have to be some kind of digital representation of currency. Essentially, instead of a $100 bill, we might imagine a $100 computer file. When Alice wants to send $100 to Bob, she attaches a $100 file to a message and sends it to him. The problem, as anyone who has sent an email attachment knows, is that sending a file does not delete it from one's computer. Alice will retain a perfect digital copy of the $100 she sends Bob, and this would allow her to spend the same $100 a second time, or indeed a third and fourth. Alice could promise to Bob that she will delete the file once he has a copy, but Bob has no way to verify this without trusting Alice.

Until recently, the only way to overcome the double spending problem was to employ a trusted third party intermediary. In our example, both Alice and Bob would have an account with a third party that they each trust, such as PayPal. Trusted intermediaries like PayPal keep a ledger of all account balances and transactions. When Alice wants to send $100 to Bob, she tells PayPal, which in turn deducts the amount from her account and adds it to Bob's. The transaction reconciles to zero. Alice cannot spend the same $100, and Bob relies on PayPal, which he trusts, to verify this. At the end of the day, all transfers among all accounts reconcile to zero. Note, however, that unlike cash, transactions that involve a third party intermediary are not final, as we have defined it, because transactions can be reversed by the third party.

In 2008, Satoshi Nakamoto (a pseudonym) announced a way to solve the double spending problem without employing third parties (Nakamoto 2008). His invention, Bitcoin, is essentially electronic cash. It allows for the first time the final transfer, not the mere copying, of digital assets in a way that can be verified by users without trusting other parties. This is accomplished through the clever use of public key cryptography, peer-to-peer networking and a proof-of-work system.

Like PayPal, the Bitcoin system employs a ledger, which is called the block chain. All transactions in the Bitcoin economy are recorded and reconciled in the block chain. However, unlike PayPal's ledger, the block chain is not maintained by a central authority. Instead, the block chain is a public document that is distributed in a peer-to-peer fashion across thousands of nodes in the

Bitcoin network. New transactions are checked against the block chain to ensure that the same bitcoins have not been previously spent, but the work of verifying new transactions is not done by any one trusted third party. Instead, the work is distributed among thousands of users who contribute their computing capacity to reconcile and maintain the block chain ledger. In essence, the whole peer-to-peer network takes the place of the one trusted third party.

Byzantine Generals Problem

Bitcoin's solution to the double spending problem – distributing the ledger among the thousands of nodes in a peer-to-peer network – presents another problem. If every node on the network has a complete copy of the ledger that they share with the peers to which they connect, how does a new node connecting to the network know that she is not being given a falsified copy of the ledger? How does an existing node know that she is not getting falsified updates to the ledger? The difficult task of reaching consensus among distributed parties who do not trust each other is another longstanding problem in the computer science literature known as the Byzantine Generals Problem, which Bitcoin also elegantly solved.

The Byzantine Generals Problem posits that a number of generals each have their armies camped outside a city that they have surrounded. The generals know that their numbers are large enough that if half their combined force attacks at the same time they will take the city, but if they do not attack at the same time they will be spread too thinly and will be defeated. They can only communicate via messenger, and they have no way of verifying the authenticity of the messages being relayed. They also suspect that some of the generals in their ranks are traitors who will send fake messages along to their peers. How can this large group come to a consensus on the time of attack without employing trust and without a central authority, especially when there will likely be attempts to confuse them with fake messages?

In essence, this is the same problem faced by Bitcoin's 'miners', the specialised nodes that verify new transactions and add them to the distributed ledger. Bitcoin's solution is to require additions to the ledger to be accompanied by the solution to a mathematical problem that is very difficult to solve but simple to verify. (This is much like calculating prime factors; costly to do, but easy to check.) New transactions are broadcast in a peer-to-peer fashion across the network by parties to those transactions. Miners look at those transactions and confirm by checking their copy of the ledger (the block chain) that they are not double-spends. If they are legitimate transactions, miners add them to a queue of new transactions that they would like to add as a new page in the ledger (a new block in the block chain). While they are doing this, they are simultaneously trying to solve a mathematical problem in which all previous blocks in the block chain are an input. The miner that successfully solves the problem broadcasts his solution to the problem along with the new block to be added to the block chain. The other miners can easily verify whether the solution to the problem is correct, and if it is they add that new block to their copy of the block chain. The process begins anew with the new block chain as an input of the problem to be solved for the next block.

The mathematical problem in question takes an average of 10 minutes to solve. This is key because the important thing is not the solution itself, but that the solution proves that the miner has expended 10 minutes of work. On average, a new block is added to the block chain every 10 minutes because the problem that miners must solve takes on average 10 minutes to solve. However, if more miners join the network, or if computing power improves, the average time between blocks will decrease. To maintain the rate at which blocks are added to six per hour, the difficulty of the problem is adjusted every 2016 blocks (every two weeks). Again, the key here is to ensure that each block takes about 10 minutes to discover.

How does this solve the Byzantine Generals Problem? Suppose that a miner is confronted with two competing block chains (just as a general might receive messages with different attack times). To choose which chain to accept and work to extend, a miner can look to see which is

longer; that is, which chain has had the most processing power devoted to it. By always choosing the longest chain, an honest miner can ensure that he is in the company of at least 51% of the other honest miners. The gap between the longest chain and competing chains will grow as time passes, since the longer chain will have more processing power behind it.

New blocks contain not just the new transactions that have been broadcast on the network, but also a transaction that assigns the winning miner 25 newly created bitcoins, which incentivises them to dedicate their computing capacity to the network. The size of the reward to miners that accompanies new blocks also halves every 210,000 blocks (every four years). The reward began at 50 bitcoins with each block when the network was launched in 2009. Today the reward is 25 bitcoins and will halve again to 12.5 in 2016. This means that the total number of bitcoins that will ever exist will not exceed 21 million. As mining rewards diminish, what incentive will miners have to lend their computing power to verify transactions? The answer is that parties to a transaction can include a transaction fee to be paid to the miner who successfully adds their transaction to a block in the block chain.

The Economics of Cryptocurrency

Governance

Cryptocurrencies do not have central banks to regulate the money supply or oversee financial institutions, but no one should neglect the importance of cryptocurrency governance institutions. We focus our discussion on two separate but interrelated ways that cryptocurrencies can be said to be governed.

Algorithmic Governance

Rules for what are considered valid cryptocurrency transactions are embedded in the peer-to-peer software that cryptocurrency miners and users run. One valid kind of transaction is the creation of new coins out of thin air. Not everyone can execute this kind of transaction – miners compete for the right to execute one of these transactions per block (on Bitcoin, every ten minutes or so). When a miner discovers a valid hash for a block, they can claim the new coins.

A transaction in which a miner claims new coins, like any other transaction, has to conform to the expectations of the network. The network will reject a block that contains a transaction in which a miner awards themselves too many new coins. The growth of coins is limited by a pre-determined amount per block.

On Bitcoin, the pre-determined amount is not scheduled to be constant over time, but rather is set to halve every 210,000 blocks, or about every four years, as described above. The total supply of bitcoins will asymptotically approach, but never exceed, 21 million. It will reach 20 million in 2025 and stop growing altogether in 2140.

Open Source Governance

The astute reader will note that the Bitcoin software that enforces particular rules about valid transactions and the rate of money creation does not appear out of thin air. Rather, the rules embedded in the software emerge from an interplay between leaders of the open source project that manages what is known as the ‘reference client’, other developers, miners, the user community and malicious actors. The dynamic between these players is as crucial to understanding Bitcoin as that of central banks, traditional monetary institutions and monetary politics is to understanding fiat currency.

Bitcoin, like all other even moderately successful cryptocurrencies to date, is a non-proprietary open source project. Users tend to look with suspicion on cryptocurrency projects that are closed source, that feature significant pre-mining in order to reward insiders, or that have other proprietary features. Other expectations of the user community also impose a check on developers. For example, the hard cap of 21 million bitcoins, while in principle subject to change through a software update, appears to be non-negotiable for Bitcoin,

although other cryptocurrencies have different money supply rules.

The division of Bitcoin software into a 'reference client' and so-called 'alt-clients' also has implications for Bitcoin's evolution. The community looks to the Bitcoin Core team for leadership as to the direction of the network. An alternative approach would be for the community to agree on the specification for the network, and then let independent teams write clients that implement the specification. The fact that Bitcoin has such a dominant reference client means that evolution can occur more quickly, although it may also have hidden costs. For example, the community has to put a lot of trust in the Bitcoin Core developers not to make bad changes to the network. A less concentrated approach to cryptocurrency development would slow down development, which would prevent any changes to the network without full deliberation of the community. It's possible that over time Bitcoin could move more to this model, but for now, the advantages of rapid evolution might outweigh the costs.

Miners also play an important role in governance. Because miners cryptographically guard against double spending, their consensus on what counts as a valid transaction is necessary for a cryptocurrency to function. A majority of miners must adopt any change to Bitcoin, and therefore the miners are able to impose a check on developers. Miners also exert influence through mining pools. Miners join pools in order to earn a more consistent payout. A single miner working alone might go for some time without discovering a block. But if miners pool their work and split their rewards, they can earn daily payouts.

Mining pools raise complications. For example, the biggest Bitcoin mining pool often has a third or more of the computing power of the Bitcoin network. If a pool ever obtained more than half of the network's computing power, it could double-spend. Double spending would destroy confidence in the Bitcoin network and would likely cause the price of bitcoins to plummet. Consequently, we observe some self-regulation by the mining pools, which are heavily invested in the success of Bitcoin. Whenever the top pool starts to approach 40% or so of computing power of the network, some participants exit the pool and join another one. So far this norm has persisted, but many in the community are concerned about mining pool concentration. Recently, the GHash.IO mining pool briefly exceeded 50 percent of Bitcoin's mining power. There is no evidence that the pool used its position to double spend, but many observers were alarmed that it was able to happen.

Concentrated mining pools have benefits as well as risks. In a crisis, it is useful to be able to assemble the key players. Such a crisis occurred on the night of 11 March 2013, when it became clear that a change in version 0.8 of the reference client introduced an unintentional incompatibility with version 0.7. As a result of the incompatibility, the two implementations of Bitcoin rejected each other's blocks, and the block chain 'forked' into two versions that did not agree on who owned which bitcoins. Within minutes of the realisation that there was a fork, the core developers gathered in a chat room and decided that the network should revert to the 0.7 rules. Over the next few hours, they were able to confer with the major mining pool operators and persuade them to switch back to 0.7, sometimes at a non-trivial cost to the miners who had mined coins on the 0.8 chain. The fact that mining pools are relatively concentrated meant that it was relatively easy to coordinate in the crisis. Within about seven hours, the 0.7 chain pulled permanently ahead and the crisis was resolved.

Another problem occurred in February 2014 when Mt. Gox, the oldest and largest Bitcoin exchange, claimed that its bitcoin holdings had been depleted through 'transaction malleability' attacks. Although it remains unclear whether Mt. Gox losses were really due to attacks, it became clear over the next several days that misunderstandings about transaction malleability were creating vulnerabilities. Some Bitcoin sites temporarily suspended withdrawals while the issues were addressed by the core development team, which updated the Bitcoin software and helped educate the community about transaction malleability, which, when properly understood, is a feature of Bitcoin, not a bug.

There is considerable scope for further study of cryptocurrency governance.

Medium of Exchange Versus Unit of Account

Bitcoin's lack of a central bank and fixed-trajectory money supply have earned it some criticism from economists concerned about macroeconomic stabilisation. Countercyclical inflationary stimulus is impossible.

However, this criticism may be misplaced. On most Keynesian and monetarist theories of monetary non-neutrality, the macroeconomic properties of money inhere in its unit-of-account function. Bitcoin is typically used as a medium of exchange without serving as a unit of account; that is, transactions will be denominated in dollars or another currency, but payment will be made using bitcoins. Unless prices, wages and contracts come to be denominated in Bitcoin, we would expect use of Bitcoin to have little cyclical impact.

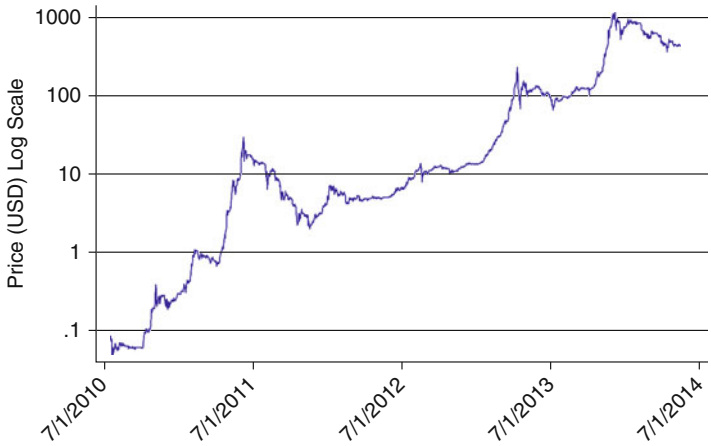
Cryptocurrencies have a number of properties that make them especially useful as media of exchange, if not as units of account. Unlike paper money, they can be transacted online as well as in person, if an Internet connection is present. Unlike credit cards, the network fee for a simple cryptocurrency transaction is low and voluntary; it is used to incentivise rapid processing of transactions by the miners. Credit card networks typically charge a swipe fee of 25 ¢ *plus* about 3% of the value of the transaction. On the Bitcoin network, transaction fees are at most a few pennies. Some retailers use merchant services to accept Bitcoin-denominated payments and have the equivalent amount of dollars deposited directly in their bank accounts. The service providers commonly charge a 1% fee for this convenience, though this may decrease as hedging costs go down (discussed below). Even with this conversion fee, merchants save 2% or more on transactions via the Bitcoin network. Another feature that could attract merchants is that customers who disavow a purchase cannot reverse most Bitcoin transactions, as they can credit card transactions.

In its separation of the medium of exchange and the unit of account, cryptocurrency brings to life some creative research from the 1970s and 1980s by economists such as Fischer Black (1970), Eugene Fama (1980), Robert Hall (1982) and Neil Wallace (1983). These authors regard the received monetary economics as highly contingent on legal and institutional arrangements; under *laissez faire*, they argue, we would observe explicit or implicit prices on media of exchange and a breakdown in the distinction between money and other financial assets. While cryptocurrency remains a niche payment mechanism and existing monetary institutions remain dominant, experimentation at the edges of our current monetary system with Bitcoin and other new cryptocurrencies could be fertile ground for new research in this tradition.

Pseudonymity and Censorship Resistance

Early news reports on Bitcoin focused on its use on the online black marketplace Silk Road. These reports propagated the misconception that Bitcoin transactions are anonymous. In fact, Bitcoin's ledger (called the block chain) is a completely public document. There is therefore a publicly accessible record of every Bitcoin transaction ever made. Bitcoin transactions occur between Bitcoin addresses, which are strings of random numbers and letters (a cryptographic hash of the address's public key). While there is no meaningful name attached to a transaction on the block chain, Bitcoin addresses function as pseudonyms for users. If a Bitcoin address can be identified as belonging to a particular individual, then all of the transactions on the block chain using that address can be attributed to that individual.

Users can take several steps to obfuscate identities and preserve some measure of financial privacy. They can generate and use a virtually unlimited number of addresses (there are 2^{160} valid Bitcoin addresses). It is considered best practice for merchants to generate a new receiving address for every transaction in order to protect their customers from scrutiny and to prevent



Cryptocurrency, Fig. 1 The price of Bitcoin (Source: Bitcoin Price Index data from CoinDesk http://www.coindesk.com/price/)

espionage from competitors. It is also becoming increasingly common for transaction processors to collate several transactions into a single one so that no one knows which address is paying which. If Alice wishes to pay Bob and Charlie wishes to pay David, a single transaction in which Alice and Charlie put in money and Bob and David take it out can make it unclear who is paying whom.

Despite the availability of these steps, the Bitcoin network remains vulnerable to sophisticated analysis. Meiklejohn et al. (2013) were able to trace bitcoins from well-known thefts through the network to centralised services such as exchanges, which in principle could be subpoenaed to reveal the identities of the criminals. They used only publicly available data; a well-equipped law enforcement agency could de-anonymise the network even further.

Although transactions are not fully anonymous, Bitcoin represents a significant shift in the enforcement burden for illegal transactions. Because non-cryptocurrency electronic payments pass through financial intermediaries, governments can enforce restrictions on transactions by regulating those intermediaries. A drug dealer cannot generally accept Visa payments because Visa will not approve a merchant whose business is dealing drugs. Illegal Bitcoin transactions may be subject to *ex post* punishment, but they are not subject to prior restraint through the regulation of financial intermediaries. This could have a significant effect on the number and kind of laws that governments are able to economically enforce.

Future developments in cryptocurrency technology could bring strong anonymity to Bitcoin or another currency. Zerocash is one proposed anonymisation system that could either be added to a future iteration of Bitcoin or released as its own currency. The strong anonymity provided by Zerocash or a similar system could have significant implications for governments who rely on controlling the financial system to enforce laws.

Pricing and Volatility

Bitcoin traded over $1 for the first time in February 2011, for $30 in June 2011, below $7 in July 2011, below $2.50 in October 2011, climbed back up to $10 by August 2012, to over $230 in April 2013, fell to below $70 within a week and rose to over $1100 in November 2013 before falling by several hundred dollars again (see Fig. 1). This volatile trend raises questions about the price of cryptocurrencies: What is the fundamental value of a Bitcoin? Why is Bitcoin so volatile? What could increase or decrease the volatility of Bitcoin in the future?

Since Bitcoin is not asset-backed, its value as a currency can only lie in its usefulness as a medium of exchange. As we have discussed, in some contexts, Bitcoin is superior to cash (e.g. it can be used online) and credit card payments (it is cheaper). In addition to its technical characteristics, its usefulness depends on the network effects that it can generate. The extent of future network

effects remains uncertain, which is perhaps the biggest reason for the volatility of Bitcoin prices so far. Some of this uncertainty will necessarily resolve itself over time, as Bitcoin is revealed either to be valueless or to have enduring value. Bitcoin is always likely to be more volatile than fiat currencies, however, because it lacks a central bank and its supply is not responsive to changes in demand.

Cryptocurrencies also raise in a new way questions of exchange rate indeterminacy. As Kareken and Wallace (1981) observed, fiat currencies are all alike: slips of paper not redeemable for anything. Under a regime of floating exchange rates and no capital controls, and assuming some version of interest rate parity holds, there are an infinity of exchange rates between any two fiat currencies that constitute an equilibrium in their model.

The question of exchange rate indeterminacy is both more and less striking between cryptocurrencies than between fiat currencies. It is less striking because there are considerably more differences between cryptocurrencies than there are between paper money. Paper money is all basically the same. Cryptocurrencies sometimes have different characteristics from each other. For example, the algorithm used as the basis for mining makes a difference – it determines how professionalised the mining pools become. Litecoin uses an algorithm that tends to make mining less concentrated. Another difference is the capability of the cryptocurrency's language for programming transactions. Ethereum is a new currency that boasts a much more robust language than Bitcoin. Zerocash is another currency that offers much stronger anonymity than Bitcoin. To the extent that cryptocurrencies differ from each other more than fiat currencies do, those differences might be able to pin down exchange rates in a model like Kareken and Wallace's.

On the other hand, exchange rate indeterminacy could be more severe among cryptocurrencies than between fiat currencies because it is easy to simply create an exact copy of an open source cryptocurrency. There are even websites on which you can create and download the software for your own cryptocurrency with a few clicks of a mouse. These currencies are exactly alike except for their names and other identifying information. Furthermore, unlike fiat currencies, they don't benefit from government acceptance or optimal currency area considerations that can tie a currency to a given territory.

Even identical currencies, however, can differ in terms of the quality of governance. Bitcoin currently has high quality governance institutions. The core developers are competent and conservative, and the mining and user communities are serious about making the currency work. An exact Bitcoin clone is likely to have a difficult time competing with Bitcoin unless it can promise similarly high-quality governance. When a crisis hits, users of identical currencies are going to want to hold the one that is mostly likely to weather the storm. Consequently, between currencies with identical technical characteristics, we think governance creates something close to a winner-take-all market. Network externalities are very strong in payment systems, and the governance question with respect to cryptocurrencies in particular compounds them.

Cryptocurrency volatility could also be reduced by the introduction of exchange-traded futures and options markets. At present, the CFTC has still not opined on the legality of cryptocurrency derivatives. However, a number of Bitcoin-based businesses have been calling for the normalisation of hedging instruments for Bitcoin, which could also have the advantage of lowering merchant processing fees. Greater access to cryptocurrency derivatives is necessary for the health of the ecosystem. Some developers have begun work on decentralised derivatives exchanges, which could be important if financial regulators refuse to approve ordinary derivatives.

Conclusion

Cryptocurrency is an impressive technical achievement, but it remains a monetary experiment. Even if cryptocurrencies survive, they may not fully displace fiat currencies. As we have tried to show in this article, they provide an interesting new perspective from which to view economic

questions surrounding currency governance, the characteristics of money, the political economy of financial intermediaries, and the nature of currency competition.

See Also

- ▶ Commodity Money
- ▶ Fiat Money
- ▶ Money

Bibliography

Black, F. 1970. Banking and interest rates in a world without money: The effects of uncontrolled banking. *Journal of Bank Research* 1: 9–20.

Fama, E.F. 1980. Banking in the theory of finance. *Journal of Monetary Economics* 6: 39–57.

Hall, R.E. 1982. Monetary trends in the United States and the United Kingdom: A review from the perspective of new developments in monetary economics. *Journal of Economic Literature* 20: 1552–1556.

Kareken, J., and N. Wallace. 1981. On the indeterminacy of equilibrium exchange rates. *Quarterly Journal of Economics* 96(2): 207–222. doi:10.2307/1882388.

Meiklejohn, S., Pomarole, M., Jordan, G., Levchenko, K., McCoy, D., Voelker, G. M., and, Savage, S. 2013. A fistful of bitcoins: Characterizing payments among men with no names. Proceedings of the 2013 Conference on Internet Measurement. http://cseweb.ucsd.edu/~smeiklejohn/files/imc13.pdf; http://dx.doi.org/10.1145/2504730.2504747

Nakamoto, S. 2008. Bitcoin: A peer-to-peer electronic cash system. bitcoin.org.https://bitcoin.org/bitcoin.pdf.

Wallace, N. 1983. A legal restrictions theory of the demand for 'money' and the role of monetary policy. *Federal Reserve Bank of Minneapolis Quarterly Review* 7: 1–7.

Cultural Transmission

Alberto Bisin and Thierry Verdier

Abstract

The economic literature analyses cultural transmission as the result of interactions between purposeful socialization decisions inside the family ('direct vertical socialization') and indirect socialization processes like social imitation and learning ('oblique and horizontal socialization'). This article reviews the main contribution of these models from theoretical and empirical perspectives. It presents the implications regarding the long-run population dynamics of cultural traits, and discusses the links with other approaches to cultural evolution in the social sciences as well as in evolutionary biology. Applications to economic problems are also briefly surveyed.

Keywords

Altruism; Cooperation; Cultural transmission; Evolutionary biology; Evolutionary economics; Genetic evolution; Identity; Imperfect empathy; Inter-generational altruism; Nature–nurture debate; Religion, economics of; Social interaction; Social norms; Socialization

JEL Classifications

I2; Z1; D9

Preferences, beliefs, and norms that govern human behaviour are partly formed as the result of genetic evolution, and partly transmitted through generations and acquired by learning and other forms of social interaction. The transmission of preferences, beliefs and norms of behaviour which is the result of social interactions across and within generations is called *cultural transmission*. Cultural transmission is therefore distinct from, but interacts with, genetic evolution.

Cultural transmission is an object of study of several social sciences, such as evolutionary anthropology, sociology, social psychology and economics, as well as of evolutionary biology. The theoretical contributions of Cavalli-Sforza and Feldman (1981) and Boyd and Richerson (1985), who apply models of evolutionary biology to the transmission of cultural traits, as well as the empirical study of cultural socialization in American schools by Coleman (1988), had a great multidisciplinary impact. Recently, economists have also studied the determination and the

dynamics of preferences, beliefs, norms and, more generally, cultural and cognitive attitudes.

Cultural transmission arguably plays an important role in the determination of many fundamental preference traits, like discounting, risk aversion and altruism. It plays a central role in the formation of cultural traits and norms, like attitudes towards the family and fertility practices, and in the job market. It is, however, the pervasive evidence of the resilience of ethnic and religious traits across generations that motivates a large fraction of the theoretical and empirical literature on cultural transmission. For instance, the fast assimilation of immigrants into a 'melting pot', which many social scientists predicted until the 1960s (see, for example, Gleason 1980, for a survey), simply did not materialize. Moreover, the persistence of 'ethnic capital' in second- and third-generation immigrants has been documented by Borjas (1992), and recently also by Fernandez and Fogli (2005) and Giuliano (2007) for norms of behaviour regarding, respectively, work and fertility practices and living arrangements. Orthodox Jewish communities in the United States constitute another example of the strong resilience of culture (see Mayer 1979, and the discussion of a 'cultural renaissance' rather than the complete assimilation of Jewish communities in New York in the 1970s). Outside the United States, Basques, Catalans, Corsicans, and Irish Catholics in Europe, Quebecois in Canada, and Jews of the diaspora have all remained strongly attached to their languages and cultural traits even through the formation of political states which did not recognize their ethnic and religious diversity.

Models of cultural transmission have implications regarding the determinants of the persistence of cultural traits and more generally regarding the population dynamics of cultural traits. In the economic literature in particular, cultural transmission is modelled as the result of purposeful socialization decisions inside the family ('direct vertical socialization') as well as of indirect socialization processes like social imitation and learning ('oblique and horizontal socialization'). Therefore, the persistence of cultural traits or, conversely, the cultural assimilation of minorities is determined by the costs and benefits of various family decisions pertaining to the socialization of children in specific socio-economic environments, which in turn determine the children's opportunities for social imitation and learning.

Evolutionary Biology Models

L. Cavalli-Sforza and M. Feldman are the first to formally study the transmission of cultural traits. Their formal models are adopted from evolutionary biology. In a baseline version of these models, they obtain a simple differential equation which describes the population dynamics of cultural traits. Consider the dynamics of a dichotomous cultural trait in the population; formally, a fraction q^i of the population has trait i, and a fraction $q^j = 1 - q^i$ has trait j. Families are composed of one parent and a child, and hence reproduction is asexual. All children are born without defined preferences or cultural traits, and are each first exposed to their parent's trait, which they adopt with probability d^i. If a child from a family with trait i is not directly socialized, which occurs with probability $1 - d^i$, he or she picks the trait of a role model chosen randomly in the population (that is, he or she picks trait i with probability q^i and trait j with probability $1 - q^i$). Therefore, the probability that the child of parents of trait i will also have trait i is $\prod^{ii} = d^i + (1 - d^i)q^i$; while the probability that he or she will have trait j is $\prod^{ij} = (1 - d^i)(1 - q^i)$. It follows that the dynamics of the fraction of the population with trait i, in the continuous time limit, are characterized by:

$$\dot{q}^i = \left(d^i - d^j\right)q^i\left(1 - q^i\right) \tag{1}$$

The dynamics that eq. (1) describes implies that the distribution of cultural traits in the population converges to a degenerate distribution concentrated on trait i whenever $d^i > d^j$ (and on trait j when $d^i < d^j$), while any initial distribution is stationary in the knife-edge case in which $d^i = d^j$. This model therefore predicts the complete assimilation of the trait with weaker direct vertical socialization. Moreover, it predicts faster

assimilation for smaller minorities. Both predictions are at odds with the documented strong resilience of cultural traits discussed above. Cavalli-Sforza and Feldman show how these extreme predictions can be relaxed by considering other effects like mutations, migrations and horizontal cultural transmission among peers. Boyd and Richerson (1985) in turn extend the analysis of Cavalli-Sforza and Feldman (1981) by considering forms of direct vertical socialization called *frequency dependent* biased transmission, which depend on the distribution of the population by cultural trait. Formally, they allow d^i to be a function of q^i.

Bisin and Verdier (2001a) study the same differential equation for the population dynamics of cultural traits, with the objective of characterizing the conditions which give rise to culturally heterogeneous stationary distributions, that is, limit population with a positive fraction of either cultural trait, $0 < q^i < 1$. They show that the crucial determinant of the composition of the stationary distribution consists in whether the socio-economic environment (oblique socialization) acts as a substitute or as a complement to direct vertical socialization. More precisely, when direct vertical socialization and oblique transmission are *cultural substitutes*, parents by definition socialize their children less the more widely dominant are their cultural traits in the population. In such a case, $d^i(q^i)$ is a strictly decreasing function in q^i, and in the long run a non-degenerate stable stationary distribution exists. It is characterized by a qi such that the direct vertical socialization of the two cultural types are equalized (that is, $d^i(q^i) = d^j(1-q^i)$): Intuitively, when family and society are substitutes in the transmission mechanism, in fact families socialize children more intensely whenever the set of cultural traits they wish to transmit is common only to a minority of the population. Conversely, families which belong to a cultural majority spend fewer resources directly socializing their children, since their children adopt or imitate with high probability the predominant cultural trait in society at large, which is the one their parents desire for them. *Cultural substitutability* tends to preserve cultural heterogeneity in the population because in this case minorities directly socialize their children more than majorities. The other typical situation is the opposite one in which direct vertical transmission is a *cultural complement* to oblique transmission; that is, when parents socialize their children more intensely the more widely dominant their cultural trait is in the population. In such a case, $d^i(q^i)$ is a strictly increasing function in qi and in the long run the dynamics converges to a culturally homogeneous cultural population (with either $q^i = 0$ or $q^i = 1$ depending on the initial distribution).

Economic Models of Cultural Transmission

Economic models of cultural transmission induce testable restrictions on the form of the function $d^i(q^i)$. In their baseline specification, for instance, Bisin and Verdier (2001a) assume that parents are altruistic towards their children and hence might want to socialize them to a specific cultural model if they think this will increase their children's welfare. If we let V^{ij} denote the utility to a type i parent of a type j child, $i, j \in \{a, b\}$, the formal assumption is

$$\text{for all } i, j \text{ with } i \neq j, V^{ii} > V^{ij}$$

This assumption, called *imperfect empathy*, can be interpreted as a form of myopic or paternalistic altruism. Parents are aware of the different traits children can adopt and are able to anticipate the socio-economic choices a child with trait i will make in his or her lifetime. However, parents can evaluate these choices only through the filter of their own subjective evaluations and cannot 'perfectly empathize' with their children. As a consequence of imperfect empathy, parents, while altruistic, tend to prefer children with their own cultural trait and hence attempt to socialize them to this trait. (Some justifications of imperfect empathy from an evolutionary perspective are provided by Bisin and Verdier 2001b. The assumption can be relaxed, as for example in Sáez-Martí and Sjogren 2005). Assume

socialization is costly and let costs be denoted by $C(d^i)$. Parents of type i then choose di to maximize:

$$-C(d^i) + \left(\prod^{ii} V^{ii} + \prod^{ij} V^{ij}\right) \tag{2}$$

$$s.t \prod^{ii} = d^i + (1 - d^i)q^i, \prod^{ij} = (1 - d^i)(1 - q^i) \tag{3}$$

Under standard assumptions, the solution to this problem provides a continuous map $d^i = d(q^i, \Delta V^i)$, where $\Delta V^i = V^{ii} - V^{ij}$ is the subjective utility gain of having a child with trait i. It reflects the degree of 'cultural intolerance' of type i's parents with respect to cultural deviations from their own trait. Given imperfect empathy on the part of parents, $\Delta V^i > 0$. The dynamics of the fraction of the population with cultural trait i is then determined by eq. (1) evaluated at $d^i(q^i) = d(q^i, \triangle V^i)$ It is straightforward to demonstrate that this class of socialization mechanisms generates cultural substitutability and therefore the preservation of cultural heterogeneity. Other micro-founded specifications and examples are provided in Bisin and Verdier (2001a), some of which illustrate the contrary possibility of cultural complementarity and the tendency of cultural homogenization over time.

Direct Socialization Mechanisms and Socio-Economic Interactions

Several specific choices contribute to direct family socialization and hence to cultural transmission. Prominent examples are education decision, family location decisions, and marriage choices While education choices have been studied by Cohen-Zada (2004), and marriage choices by Bisin and Verdier (2000), the literature has to date shown little interest in the socialization effects of location choices, for instance, the socialization effects of urban agglomeration by ethnic or religious trait.

The simple analysis of the economic model of cultural transmission of Bisin and Verdier depends crucially on the assumption that the utility to a type i parent of a type j child, V^{ij} is independent of the distribution of the population by cultural trait, that is, independent of q^i. Many interesting analyses of cultural transmission require this assumption to be relaxed. In many instances the adoption of the cultural trait of the majority in fact favours children, for example in the labour market; a typical example is language adoption. In this case altruistic parents, even if paternalistic, might favour (or discourage less intensely) the cultural assimilation of their children. If we allow for interesting socio-economic effects interacting with the socialization choices of parents, the basic cultural transmission model of Bisin and Verdier has been applied to several different environments and cultural traits and social norms of behaviour, from preferences for social status (Bisin and Verdier 1998) to corruption (Hauk and Sáez-Martí 2002), hold-up problems (Olcina and Penarrubia 2004), development and social capital (François 2002), intergenerational altruism (Jellal and Wolff 2002), labour market discrimination (Sáez-Martí and Zenou 2005), globalization and cultural identities (Olivier et al. 2005), and work ethics (Bisin and Verdier 2005).

Empirical Analysis of Cultural Transmission Models

While an interesting literature has documented the relevance of cultural factors in several socio-economic choices, much less is known about cultural transmission per se. Nonetheless, several important questions are beginning to be answered. First of all, several important correlations have been documented in sociology, in particular with regard to the role of marriage in socialization (see, for instance, Hayes and Pittelkow 1993; Ozorak 1989; Heaton 1986). The literature in economics has instead concentrated more specifically on the direct empirical validation of the economic approach to cultural transmission surveyed above, thereby estimating the relative importance of direct and oblique socialization for different specific traits and the prevalence of cultural

substitution or complementarity in specific socio-economic environments. Patacchini and Zenou (2004) find evidence of cultural complementarity in education in the United Kingdom. Cohen-Zada (2004) finds instead for the United States that the demand for private religious schooling decreases with the share of the religious minority in the population, in accord with cultural substitution. Fernandez et al. (2004) find evidence of an important role for mothers in the transmission to their sons of attitudes favouring the participation of women in the labour force and acquisition of higher education. Finally, Bisin et al. (2004a), using the General Social Survey data for the United States over the period 1972–96, estimate for religious traits the structural parameters of the model of marriage and child socialization in Bisin and Verdier (2000). They find that observed intermarriage and socialization rates are consistent with Protestants, Catholics and Jews having a strong preference for children who identify with their own religious beliefs, and taking costly decisions to influence their children's religious beliefs. The estimated 'relative intolerance' parameters are high and asymmetric across religious traits, suggesting an interestingly rich representation of 'cultural distance'.

Genetic and Cultural Evolution

Cultural transmission possibly has a role also in the determination of fundamental preference parameters, such as time discounting, risk aversion, altruism, and interdependent preferences. Purely evolutionary models have been complemented by alternative models of cultural transmission and genetic and cultural co-evolution. The wealth of different approaches proposed is best exemplified by the study of preferences for cooperation. The observation that humans often adhere to collectively beneficial actions which are not in their private interest (or which are not rationalizable as strategic equilibria) has led to a theoretical literature explaining how psychological 'preferences for cooperation' can be sustained in the context of genetic and/or cultural evolution (this is called the *puzzle of pro-sociality* by Gintis 2003a). For instance, in the context of the Prisoner's Dilemma, Becker and Madrigal (1995) exploit the ability of habits to induce preferences; Guttman (2003), Stark (1995), and Bisin et al. (2004b) show how cooperation can be sustained by different modes of cultural evolution; Gintis (2003b) shows that a general capacity to internalize fitness-enhancing norms of behaviour can be genetically adaptive, and hence that cooperation can also be internalized by 'hitchhiking' on this general capacity.

The empirical evidence on the nature–nurture debate (see Ceci and Williams 1999, for a review) has not yet been systematically taken to the point of distinguishing the genetic from the cultural factors in the determination of fundamental preference parameters. Similarly, the empirical evidence distinguishing the different cultural transmission models of fundamental preference traits is almost non-existent. The only exception is by Jellal and Wolff (2002), who study the implication of the pattern of *inter vivos* transfers within the family in France for the transmission of intergenerational altruism. They argue that the evidence is more consistent with a cultural transmission model such as that of Bisin and Verdier (2001a) rather than with a 'demonstration effect' model, as in Stark (1995), where parents take care of their elders in order to elicit similar behaviour in their children.

See Also

- ▶ Culture and Economics
- ▶ Identity
- ▶ Social Interactions (Empirics)

Bibliography

Becker, G., and V. Madrigal. 1995. *On cooperation and addiction*. Mimeo: University of Chicago.

Bisin, A., G. Topa, and T. Verdier. 2004a. Religious intermarriage and socialization in the United States. *Journal of Political Economy* 112: 615–664.

Bisin, A., G. Topa, and T. Verdier. 2004b. Cooperation as a transmitted cultural trait. *Rationality and Society* 16: 477–507.

Bisin, A., and T. Verdier. 1998. On the cultural transmission of preferences for social status. *Journal of Public Economics* 70: 75–97.
Bisin, A., and T. Verdier. 2000. Beyond the melting pot: Cultural transmission, marriage and the evolution of ethnic and religious traits. *Quarterly Journal of Economics* 115: 955–988.
Bisin, A., and T. Verdier. 2001a. The economics of cultural transmission and the dynamics of preferences. *Journal of Economic Theory* 97: 298–319.
Bisin, A., and T. Verdier. 2001b. Agents with imperfect empathy might survive natural selection. *Economics Letters* 2: 277–285.
Bisin, A., and T. Verdier. 2005. *Work ethic and redistribution: a cultural transmission model of the welfare state*. Mimeo: New York University.
Borjas, G. 1992. Ethnic capital and intergenerational income mobility. *Quarterly Journal of Economics* 57: 123–150.
Boyd, R., and P. Richerson. 1985. *Culture and the evolutionary process*. Chicago: University of Chicago Press.
Cavalli-Sforza, L., and M. Feldman. 1981. *Cultural transmission and evolution: A quantitative approach*. Princeton: Princeton University Press.
Ceci, S., and W. Williams. 1999. *The nature–nurture debate: The essential readings*. Oxford: Blackwell.
Cohen-Zada, D. 2004. *Preserving religious identity through education: economic analysis and evidence from the US*. Mimeo: Ben-Gurion University.
Coleman, J. 1988. Social capital in the creation of human capital. *American Journal of Sociology* 94: S95–S120.
Fernandez, R., and A. Fogli. 2005. Culture: An empirical investigation of beliefs, work, and fertility. Working paper no. 11268. Cambridge, MA: NBER.
Fernandez, R., A. Fogli, and C. Olivetti. 2004. Mothers and sons: Preference formation and female labor force dynamics. *Quarterly Journal of Economics* 119: 1249–1299.
François, P. 2002. *Social capital and economic development*. New York: Routledge.
Gintis, H. 2003a. Solving the puzzle of prosociality. *Rationality and Society* 15: 155–187.
Gintis, H. 2003b. The hitchhikers guide to altruism: genes, culture and the internalization of norms. *Journal of Theoretical Biology* 220: 407–418.
Giuliano, P. 2006. Living arrangements in Western Europe: does cultural origin matter? *Journal of the European Economic Association* 5: 927–952.
Gleason, P. 1980. American identity and Americanization. In *Harvard Encyclopedia of American Ethnic Groups*, ed. T. Stephan, O. Ann, and H. Oscar. Cambridge, MA: Harvard University Press.
Guttman, J. 2003. Repeated interaction and the evolution of preferences for reciprocity. *Economic Journal* 113: 631–656.
Hauk, E., and M. Sáez-Martí. 2002. On the cultural transmission of corruption. *Journal of Economic Theory* 107: 311–335.
Hayes, B., and Y. Pittelkow. 1993. Religious belief, transmission, and the family: An Australian study. *Journal of Marriage and the Family* 55: 755–766.
Heaton, T. 1986. How does religion influence fertility? The case of Mormons. *Journal for the Scientific Study of Religion* 28: 283–299.
Jellal, M., and F. Wolff. 2002. Cultural evolutionary altruism: Theory and evidence. *European Journal of Political Economy* 18: 241–262.
Mayer, E. 1979. *From suburb to shetl: The jews of boro park*. Philadelphia: Temple University Press.
Olcina, G., and C. Penarrubia. 2004. Hold-up and intergenerational transmission of preferences. *Journal of Economic Behavior and Organization* 54: 111–132.
Olivier, J., M. Thoenig, and T. Verdier. 2005. *Globalization and the dynamics of cultural identity. Mimeo*. Paris: Paris-Jourdan Sciences Économiques.
Ozorak, E. 1989. Social and cognitive influences on the development of religious beliefs and commitment in adolescence. *Journal for the Scientific Study of Religion* 28: 448–463.
Patacchini, E., and Y. Zenou. 2004. Intergenerational education transmission: neighborhood quality and/or parents' involvement? Working paper no. 631. Stockholm: Research Institute of Industrial Economics.
Sáez-Martí, M., and A. Sjogren. 2005. Peers and culture. Working paper no. 642. Stockholm: Research Institute of Industrial Economics.
Sáez-Martí, M., and Y. Zenou. 2005. *Cultural transmission and discrimination. Mimeo*. Stockholm: Research Institute of Industrial Economics.
Stark, O. 1995. *Altruism and beyond: An economic analysis of transfers and exchanges within families and groups*. Cambridge: Cambridge University Press.

Culture and Economics

Raquel Fernández

Abstract

Modern neoclassical economics has, until recently, ignored the potential role of culture in explaining variation in economic outcomes, largely because of the difficulty in rigorously separating the effects of culture from those of institutions and traditional economic variables. This article selectively reviews some recent attempts to empirically identify the effects of culture on economic outcomes and to answer the question, 'does culture matter and, if so,

how much?' Open theoretical and empirical questions are discussed, including the relationship between culture and institutions.

Keywords

Agency problems; Assortative matching; Cultural assimilation; Culture and economics; Endogenous preferences; Epidemiology; Fertility; Human capital; Institutions; Instrumental variables; International migration; Multiple equilibria; Occupational selection; Religion; Social norms; Technology; Trust; Women's work and wages; Beliefs; Identity; Preference transmission

JEL Classifications

O4; Z1

Economic decisions are made within a social context; as Aristotle reminds us, man is a social animal. The relevance of this statement to economics, however, is far from clear. In what ways, if any, do we need to consider the social nature of man in order to study economic questions? This article attempts to provide a partial answer to this question.

Traditionally, economists seek to explain differences in economic outcomes by studying how agents, with given preferences and beliefs, react to changes in the policy environment, institutions and technology. At a deeper level than the taste for apples versus oranges, however, few would deny that preferences and beliefs must be, to some extent, endogenous. Our level of trust in others, the determinants of status in society, our beliefs about the correct trade-off between efficiency and equity, or the 'proper' roles for men and women, are all examples of beliefs or preferences that have differed across societies and over time. These beliefs and preferences impact on individual behaviour and how society allocates scarce resources. At the individual level they help determine whether a woman participates in the formal labour market and the career she follows, the extent to which racism is tolerated, or the degree of assortative matching on wealth in marriages. At a collective level, they help determine, for example, the range and depth of the welfare state, the legality of slavery, or the proportion of the budget that is dedicated to foreign aid.

Although at some general level few may disagree that preferences, beliefs, or values of the type discussed above are endogenous (and may therefore differ across societies), whether they have a quantitatively significant impact on economic outcomes is another matter. Do differences in beliefs and preferences that vary systematically across groups of individuals separated by space (either geographic or social) or time – what I shall henceforth term *culture* – play an important role in explaining differences in outcomes? (For the purposes of this article, I will not give a more rigorous definition of culture than the abbreviated one here. See Elster 1989, for a discussion of social norms and culture and Manski 2000, for a discussion of peer effects and social interactions.) Modern economics (as opposed to sociology or anthropology) has largely been, until recently, reluctant to investigate this question. Although in principle there is nothing non-standard about positing preference/belief heterogeneity among individuals to explain differences in outcomes, the Stigler–Becker dictum *de gustibus non est disputandum* (Stigler and Becker 1977) and its assertion that 'no scientific behavior has been illuminated by assumptions of differences in taste' has cast a long shadow in economics. Thus, the main challenge faced by those who believe that culture might matter has been to find a convincing way to show that culture can be studied rigorously and, in particular, that it is possible to separate the influence of culture from institutions and standard economic variables. In this sense, running, say, cross-country regressions on variables that one suspects reflect cultural attitudes (for example, different savings patterns may reflect attitudes towards thrift) to study the effect of culture has long (and correctly) been considered unsatisfactory. Despite one's best efforts to control for differences in countries' economic environments, identifying the residual with culture is ultimately unconvincing. It is difficult, if not impossible, to summarize the economic environment faced by agents with a few aggregate variables. Thus, there are bound to be omitted

variables and problems of endogeneity, which are all further confounded by mismeasurement.

Hence, despite a long history of writers on the relationship between culture and economics (which includes Marx, Weber, Gramsci, Polanyi, Banfield and, more recently, Putnam and Landes, among others), modern neoclassical economics has been by and large silent on the topic of culture and only in recent years have economists started to think seriously again about how culture may help explain economic phenomena. In this article I will selectively review some recent attempts to empirically identify the effects of culture on important economic outcomes and to answer the question, 'does culture matter?' Answering this question affirmatively naturally leads one to explore the propagation mechanisms of culture, to theorize about the relationship between institutions and culture, and to investigate the dynamic of culture – all topics that I will briefly touch upon at the end.

Empirical Evidence on Culture

In this section I examine some of the recent evidence on the importance of culture for economic outcomes. For expository ease, I have divided the empirical evidence into that which uses survey data, evidence based on immigrants or their descendants (what I call the 'epidemiological approach'), and historical case studies. There is also a small body of experimental work that, by showing that across societies there exist marked differences in how individuals play games such as the ultimatum, public good or dictator game, has also shed light on the relationship between culture and economics (see, for example, Henrich et al. 2001).

Survey-based Evidence

Perhaps the most natural approach to doing empirical work on culture consists in using the beliefs expressed by individuals in surveys (for instance, the World Value Surveys) on a variety of issues as expressions of culture and correlating them with economic outcomes. This approach, however, must overcome the problem of reverse causality. That is, differences in beliefs may be solely a consequence of different economic and institutional environments. Hence, the use of instrumental variables is required in order to identify causality. Overall, this has been difficult to achieve.

As shown by Guiso et al. (2003), the intensity of religious beliefs and religious denomination are correlated with a variety of individual attitudes such as trust in others, government's role, views of working women and the importance of thrift. Guiso et al. (2006) show that these attitudes, aggregated at the country level, are correlated with cross-country aggregate outcomes (for example, savings, redistributive versus regressive taxation, and trade). In order to ensure that the reverse causality is not at play, the attitudes are instrumented, usually by the religious composition in the country. This work is suggestive but there are several concerns associated with it. In addition to questions about omitted variables, it is not clear that religious composition is a valid instrument since it may also help explain the aggregate outcome through other channels. (Indeed, the coefficients on the instrumental variable results tend to look very high relative to the ones obtained by ordinary least squares. Running regressions at the individual outcome level would be more convincing, but opinion surveys unfortunately tend not to have high-quality economic data (the World Value Survey, for example, classifies income levels into ten categories). Recent work by Guiso et al. (2005) on the relationship between trust and trade, instead instruments trust with the genetic distance between indigenous populations. This seems a promising avenue of research.

Tabellini (2005) takes a significant step towards overcoming some of the weaknesses discussed above. To study whether culture affects economic development across European regions, he also aggregates (at the regional level) individual responses from the World Value Surveys to questions about trust, respect and the link between individual effort and economic success. The scope for omitted variables is reduced by focusing on within-country variation in Europe (by including country fixed effects). The attitudes are then

instrumented with historical variables, such as regional literacy rates at the end of the 19th century and indicators of political institutions in the period from 1600 to 1850. The author finds that the proxies for culture are quantitatively significant determinants of per capita GDP levels and growth rates across regions. It is possible of course that the instruments are not valid. For example, they could affect output directly via sectoral composition or public investment. The paper contains a good discussion of these and other alternative hypotheses.

The Epidemiological Approach

A very different approach to relying on opinion data is to examine the economic outcomes of immigrants or their descendants. This is reminiscent of the epidemiology literature that, in order to attempt to identify the contribution of the environment broadly defined (namely, physical and cultural) relative to genes in disease, studies various health outcomes for immigrants and compares them to outcomes for natives (see, for example, the classic study by Marmot et al. 1975).

To understand the strengths and weaknesses of such an approach, suppose that the level of, say, heart disease differs markedly between two countries (the source and host countries). If heart disease in immigrants converges to that of natives in the host country, the difference between the two countries is unlikely to be driven by genetics and instead results from the environment. Failure to find convergence, on the other hand, does not imply the opposite. There are many reasons why the environment may be solely responsible and still sustain differential levels of heart disease. For example, cultural assimilation may occur slowly (for instance, if immigrants maintain the same dietary patterns as in the source country), or living in the source country at a young age may confer some degree of immunity, or selection into immigration may be correlated with a particular health outcome.

The epidemiological strategy in economics has its own set of problems. In particular, it is important to recognize that immigrants may be subject to many shocks (language difficulties, worse employment opportunities, greater uncertainty and so forth) which cause them to deviate from their traditional behaviour. Culture, furthermore, is socially constructed: to be replicated, the behaviour may require the incentives – rewards and punishments – provided by a larger social body such as a neighbourhood, school, or ethnic network. Furthermore, immigrants are unlikely to be a representative sample of their home-country's population. Their beliefs, preferences, and unobserved differences in their economic circumstances may differ significantly from the country average. Lastly, the exposure of immigrants (or their descendants) to a different culture from the one prevalent in their country of heritage presumably weakens the latter's impact on their behaviour. Note that all the factors mentioned above introduce a bias towards finding culture to be insignificant. Thus, on the whole, comparisons of behaviour or outcomes across different immigrant groups are a very demanding test of the importance of culture. In epidemiology, when differences across groups remain, one must be careful not to conclude that genetics is determinative when the underlying cause may be cultural; in economics, when significant differences are not observed, one must be careful not to rule out cultural forces.

In economics, the paper by Carroll et al. (1994) is the first that, to my knowledge, follows an approach similar to the one described above. The authors are interested in exploring whether cross-country differences in savings rates may be culturally driven. Using individual-level data on immigrants to Canada, they estimate individual consumption levels as a function of permanent income (as captured by labour and asset income), the interaction of this variable with demographic variables, some measures of wealth, and finally the interaction of a region of origin dummy (and years since arrival to Canada) with their measure of permanent income. If there exist different cultural attitudes towards savings, and if this attitude is maintained in immigrants, then one should observe different propensities across immigrants, by region of origin, to consume out of permanent income (that is, the regional dummies should be significantly different from one another). The authors find that the saving patterns of immigrants

do not vary significantly by region of origin. Recent immigrants as a whole save less than native-born Canadians, but there is no statistically significant difference in behaviour across immigrant groups.

There are several weaknesses in the data-set used in the study above that may bias it against finding results that show a significant impact of culture. Wealth, for example, is not well measured. In particular, as only South East Asia's saving rate differed markedly from those of other regions in the immigrant population (31 per cent relative to 18–20 per cent across the remaining regions), the small number of immigrants from this group in the sample limits the power of the test. Note also that, if the motivation to save more stems from the desire to provide one's child with greater status via a larger bequest, the incentive to do this may be much less marked in a society in which savings are generally low or in which status stems from consumption behaviour.

Fernández and Fogli (2005, 2006) use a similar, but arguably less problematic, methodology by studying second-generation Americans in order to investigate the quantitative importance of culture. Their research focuses on the fertility and work behaviour of married second-generation American women (that is, women who were born in the United States but whose parents were born elsewhere). The use of second-generation immigrants attenuates the problems associated with the first generation's adjustment to a foreign setting (for example, language difficulties) and even some selection problems are less likely to play a role for the second generation. On the other hand, second-generation individuals have been more exposed to the new culture, and that will tend to diminish the role of culture from the country of heritage. Our hypothesis is that attitudes towards woman's 'proper' role in society and towards ideal family size are culturally different across countries and that this culture is likely to be transmitted intergenerationally and show up in systematic differences in female labour force participation (LFP) and fertility, even if individuals were raised in the United States.

In our 2005 paper, the challenge was how to best capture the attitudes towards women and family size in the parents' country of origin. We chose not to use country dummies (as in Carroll et al. 1994) but to instead examine whether past values of economic variables in the country of origin that should reflect this culture – in particular, past values of female LFP and total fertility rates (TFR) – are able to play a quantitatively significant role in explaining differences in outcomes across second-generation women in the United States. Our argument is that these economic variables reflect the institutions (for example, markets, legal framework, minimum wages and so on), the strictly economic environment (demand and supply, transportation costs, access to day care, for example), as well as the preferences and beliefs (that is, the culture) of individuals in the country making decisions at that time. If these variables are able to explain the behaviour of women who, by virtue of living in the USA and in a different time period, face different institutions and economic variables, then solely the cultural component of these variables should affect their choices. This is a more demanding test that is superior to the 'black box' approach of using country dummies which leaves open the question of what it is about the country that matters to outcomes.

In individual level regressions, we find that our cultural proxies – past values of female LFP and TFR – help explain both how much second-generation American women work and their fertility. As our data-set – the 1970 US Census – does not allow us to control for family factors such as parental wealth, income, and education, we include the woman's education, her spouse's education, and total personal income (as well as location, age, and so on) in our regressions. By including these variables, the coefficient on the cultural proxy only captures the direct effect of culture rather than its full direct and indirect effects (for example, a woman who wants engage in market work is more likely to invest in education and hence, by controlling for education, we are eliminating the effect of culture on this variable), but this is preferable to not controlling for differences in parental background, other than culture, that may affect women's work and fertility outcomes. We find that the cultural proxies still

matter even after including these additional variables. Furthermore, the cultural proxies are quantitatively significant: a one standard-deviation increase in the corresponding cultural proxy is associated with approximately an eight per cent increase in hours worked per week and about a 14 per cent increase in the number of children. The forces of assimilation means that these numbers should be taken, if anything, as a downward biased estimate of the true power of culture in the original setting (that is, in the country of ancestry).

We also examine the most compelling alternative economic explanation for our results, namely, the hypothesis that these are driven by unobserved human capital. We do this by showing that the results are robust to the inclusion of the country of ancestry's level of per capita GDP in various years and to the years of education of immigrants (by country of ancestry) in 1940 (this remains the case when Hanushek and Kimko's (2000) measures of education quality in the parents' country of origin are included). We also demonstrate that the work cultural proxy does not have explanatory power in a Mincer wage regression which it would be expected to have if it captured unobserved human capital. Lastly, we show that the work cultural proxy is insignificant in explaining how much married second-generation American men work whereas the fertility cultural proxy retains its explanatory power. (If the work cultural proxy had a negative effect on how much these men work, that might indicate a substitution effect. In our regressions, the coefficient is basically zero and insignificant.) This is important because it implies that there does not exist some omitted economic variable at the parental country-of-origin level that affects the productivity of both men and women and that helps explain how much they work.

The methods used in Fernández and Fogli (2005) could be profitably extended to examine other issues, such as entrepreneurship or savings behaviour. It might also be interesting to elaborate upon the recent approach by Algan and Cahuc (2006) that attempts to combine survey evidence with the epidemiological approach in order to study the effects of culture on cross-country labour market outcomes. Although this work is too preliminary to discuss in depth, using the attitudes of, say, second-generation Americans to instrument for the attitudes of individuals in the home country seems cleaner than relying on variation in religious denominations. As usual, the question will be whether there is some omitted background economic variable correlated with the country of origin (particularly given the quality of the survey data-sets) that could be driving the results, but it seems a promising avenue of research (see also the interesting work on culture and migrants within regions in Italy by Ichino and Maggi 2000. As shown recently in Fernández (2007a) using the World Value Survey, the attitudes of individuals in the country of ancestry towards women's market work and housework have explanatory power for the work outcomes of second-generation American women in 1970.

Historical Case Studies

The analysis of historical episodes in which changes in either culture or environment yield 'natural experiments' is likely to add richness and depth to our understanding of culture and the economy. Greif's 1994 paper is probably the best-known work in economics that makes the link between culture and institutional development. In brief, Greif argues that cultural beliefs (collectivist versus individualist) are reflected in the different ways in which in the 11th century Genoese traders and Maghrebi traders set up their trading institutions. Both groups of merchants required agents to conduct their business overseas, and in both cases there was an agency problem as the overseas agent might be tempted to cheat the merchant. Maghrebi traders set up 'horizontal' relations in which merchants served as agents for traders and vice versa. Information was shared among merchants/traders and an agent who was dishonest with one merchant could expect to be shunned by other merchants. The Genoese, on the other hand, set up 'vertical' relationships in which individuals specialized as merchants or agents. Information was not shared among merchants. This led the Genoese to set up more formal enforcement institutions. The two different responses, argues Greif, then had important consequences once trading opportunities

were expanded in previously inaccessible areas. The Maghrebi expanded trade using other Maghrebi agents whereas the Genoese were able to establish agency relations with non-Genoese, leading to very different economic development paths thereafter (see also Greif's 2005, recent book on the topic).

Another compelling example is provided by Botticini and Eckstein (2005) who present the thesis that an 'exogenous' cultural change gave rise to the pattern of Jewish occupational selection that we see to this day. They argue that with the destruction of the Temple in Jerusalem in 70 CE, the Pharisees became the dominant religious group and transformed Judaism from a religion based on sacrifices to one whose main rule required each male to read and to teach his sons the Torah. This reform was implemented in places where most Jews were farmers who would not gain anything from investing in education. When urbanization expanded many centuries later, Jews had a comparative advantage in the skilled occupations demanded in the new urban centres. Thus, culture – the religious requirement of reading skills for other than human capital reasons – gave rise to the pattern of Jewish occupational selection seen since the ninth century.

Theories of Culture

Is it necessary to modify the standard economic model in order to incorporate culture? The answer definitely is 'no'. What appear to be societal differences in preferences may only be choice of equilibrium strategies in a game with multiple equilibria and standard preferences. This is in fact the most common way to think about the role of culture in economics, and is fully in keeping with our working definition of culture as systematic differences (across groups) in preferences or beliefs. Here the heterogeneity lies in the expectations (beliefs) over the strategies that will be played in equilibrium. Hence differences in culture can be identified with, for example, which equilibrium we play in a static game (for example, do we drive on the right- or left-hand side of the road) or the degree of cooperation ('trust') sustained in a repeated Prisoner's Dilemma game.

Within the 'culture as multiple equilibria' literature, I find particularly interesting the research that attempts to generate behaviour that looks like social norms (such as determinants of status). Take, for example, a dynamic matching model in which individuals who differ in wealth choose a partner with whom to match and obtain utility from joint consumption and the utility of their child. As shown in Mailath and Postlewaite (2003), in addition to an equilibrium in which there is assortative matching on wealth, there may also be an equilibrium with imperfectly assortative matching that depends also on non-economic characteristics such as whether one has blue eyes. In this equilibrium, blue eyes matter not because of their intrinsic value, but simply because the matching rule allocates, for the same wealth level, a wealthier partner to individuals with blue eyes. Thus, a woman would be willing to match with a man with blue eyes and slightly lower wealth than another man without blue eyes, because although she obtains lower joint consumption, there is a 50 per cent chance that her child would inherit blue eyes and hence a better match and higher consumption in the future. To an outside observer, it might therefore appear that in this society people had an intrinsic preference for blue eyes, although this inference would be incorrect.

Although the example above is interesting, its explanation for a particular social norm seems incomplete and intuitively less than compelling. The preference for blue eyes or light skin may perhaps initially come about as a choice among many equilibria and involve solely a calculation about the trade-off between one's own consumption and that of one's child (though that too seems doubtful and is more likely the result of a history in which these traits are correlated with higher status). Over the longer run, however, one may conjecture that what sustains these equilibria – what makes these cultural traits less fragile to perturbations – is that these calculations are embodied in the individual and in society as preferences and beliefs about the inherent superiority/desirability of such features. People come to

prefer blue eyes; people become racist. Thus, what is missing more generally in the theory of culture is an analysis of how preferences and beliefs (about things other than equilibrium strategies) themselves evolve.

The hypothesis that certain features of culture (those that have greater depth than driving on the left or the right side of the street) become part of preferences and beliefs implies that they cannot be discarded easily simply because they are no longer useful or beneficial, though over time this will certainly lessen their appeal. In this way, the operation of culture may be clearest to perceive when it no longer serves any useful societal purpose or particular group interest but nonetheless, at least for some time, persists – for example, religious prohibition on eating pork. (One reason speculated for this prohibition is that consumption of undercooked pork is linked to trichinosis. It is now known that this problem can be eliminated, however, by thoroughly cooking the meat.) In the context of the matching example above, individuals may eventually be willing to match with lower wealth people with blue eyes because this matching rule is incorporated into preferences/beliefs over what type of mate is intrinsically better even if the benefit derived by passing this trait on to their offspring is no longer substantial (say, because family size falls and decreases the payoff from the inheritable trait relative to the decrease in immediate joint consumption).

So far, we have discussed differences in culture as systematic differences in preferences and beliefs without distinguishing much between the two. This is not accidental, since, in general, the distinction between preferences and beliefs for our purposes is rather fuzzy. Even for simple preferences such as the trade-off between apples and oranges, what one knows (or believes) about the nutritional contents of the two may affect how one 'feels' about them, as may any other mental associations (for example, whether one is considered more exotic, how they were grown and so forth). In general, there are few pure (or naive) preferences – what one thinks or believes influences how one feels (and the same may be true vice versa. See Damasio 1995, for an interesting exposition of evidence in favour of the hypothesis that emotions affect – and in fact are necessary for – the ability to think well). This is not to deny that people have some inherent tastes (for example, it is believed that human beings have a taste for fat, probably because of the evolutionary advantage associated with an inclination to eat meat in an environment in which protein and iron were not easily obtained).

For more complex questions the above is even more likely to be true. Consider, for example, the large increase in female labour force participation in the 20th century. Is it that woman's disutility from market work decreased or that her beliefs about the meaning or consequences of her working that changed over time? The dichotomy between the two alternatives does not seem very useful in this case. If the focus is on understanding why actions change over time, then using standard preferences and modelling the evolution of beliefs as giving rise to changes in expected payoffs may be the more useful strategy (the latter is the approach taken by Fernández 2007b, who shows that a model of the evolution of female LFP as an intergenerational learning process does a good job of replicating a century of US female LFP data). If instead one wished to understand the utility from a given action, particularly one in which identity is concerned, then incorporating cultural beliefs into preferences may be a better route (see, for instance, Akerlof and Kranton 2000). For example, wearing a dress or having a woman as a boss may decrease a man's utility, independently of any expectations of future consequences, simply because it makes him feel (culturally) less masculine.

Culture and Institutions

As seen previously, the main challenge faced by most empirical work on culture is to convincingly isolate its effects from the incentives provided by traditional economic variables and institutions. This should not be taken to mean that culture and institutions are independent variables. Indeed, one way to think about institutions is as congealed culture: that is, which institutions are set up and how these evolve depends not only on the problems faced by society (or by a particular group in society) at a particular moment in time but also the

beliefs/preferences – the culture – that are prevalent. As elaborated on in our earlier discussion of Greif (1994), cultural beliefs (collectivist versus individualist), for example, were reflected in the different ways in which in the 11th century Genoese traders and Maghrebi traders set up their trading institutions, leading to very different economic development paths thereafter. My hypothesis is that the reverse causality is also likely to hold: that is, not only does culture affect institutions but also institutions affect the dynamic evolution of culture. In this sense, work that attempts to establish whether institutions or culture are the most important determinants of economic development seems misconceived (see Fernández 2007c, for a theoretical analysis of the dynamic dependency of culture and institutions; also Bowles 1998, for a review of some of the theoretical and empirical evidence on the effect of markets on culture).

Concluding Remarks

The rigorous study of culture and economics is in its infancy. We would like to understand, for example, how culture propagates and evolves. In particular, what is the relative importance of family versus other institutions as cultural transmission mechanisms for different beliefs or in different environments? To what extent is cultural transmission purposeful, that is, optimizing on the part of an individual or her parents (as in Bisin and Verdier 2000) or for a social group, and to what extent is it involuntary? (Fernández et al. 2004, show that whether a man's mother worked while he was growing up is correlated with whether his wife works, even after controlling for a whole series of socioeconomic variables. They interpret this as preference transmission, but whether it is voluntary – optimizing – or simply by example is an open question.) When and why does culture change abruptly whereas at other times it proceeds glacially?

The relationship between technology and culture also needs to be investigated. How does technology influence culture and how does culture shape technological change? Some papers (for instance, Greenwood and Guner 2005; Greenwood et al. 2002) argue that sexual norms and female LFP changed because of changes in technology. These papers ignore, among other things, the endogeneity of demand for new technology. Despite the convenient simplification of treating technology as a primitive, it too is endogenous. The extent to which societies put resources into developing technology that 'liberates' individuals from household work, for example, depends on things such as whether slavery is available or whether women expect to work in the market or at home. Put differently, both the relative price of market versus household labour and the elasticity of labour supply depend on the institutions (for example, slavery) and expected division of labour (for example, clearly differentiated gender roles) that are in place. The opposite is also true – the extent to which one can substitute capital for labour, whether at work or at home, helps determine which institutions are viable and may determine the pace and ease with which beliefs or preferences change.

From a theoretical perspective, the endogeneity of preferences and beliefs raises difficult questions for welfare. How should we evaluate policies once we recognize that preferences can change? While this is indeed a vexing and problematic question for welfare economics, recognizing that man is a social animal that is (perhaps uniquely) capable of reflecting upon, and hence changing, his preferences and beliefs greatly enriches our view of ourselves and the world and within it the potential role of economic discourse. In the words of A.O. Hirschman, *'de valoribus est disputandum'*.

See Also

- ▶ Cultural Transmission
- ▶ Social Norms

Bibliography

Algan, Y., and P. Cahuc. 2006. *Minimum wage: The price of distrust*. Mimeo, CREST-INSEE.

Akerlof, G., and R.E. Kranton. 2000. Economics and identity. *Quarterly Journal of Economics* 115: 715–733.

Bisin, A., and T. Verdier. 2000. Beyond the melting pot: Cultural transmission, marriage, and the evolution of ethnic and religious traits. *Quarterly Journal of Economics* 115: 955–988.
Botticini, M., and Z. Eckstein. 2005. Jewish occupational selection: Education, restrictions, or minorities? *Journal of Economic History* 65: 922–948.
Bowles, S. 1998. Endogenous preferences: The cultural consequences of markets and other economic institutions. *Journal of Economic Literature* 36: 75–111.
Carroll, C., B. Rhee, and C. Rhee. 1994. Are there cultural effects on saving? Some cross-sectional evidence. *Quarterly Journal of Economics* 109: 685–699.
Damasio, A. 1995. *Descartes' error: Emotion, reason, and the human brain*. New York: Harper Perennial.
Elster, J. 1989. Social norms and economic theory. *Journal of Economic Perspectives* 3(4): 99–117.
Fernández, R. 2007a. Women, work, and culture. *Journal of the European Economic Association* 5: 305–332.
Fernández, R. 2007b. *Culture as learning: The evolution of female labor force participation over a century*. Mimeo, New York University.
Fernández, R. 2007c. *The co-evolution of culture and institutions*. Mimeo, New York University.
Fernández, R., and A. Fogli. 2005. *Culture: An empirical investigation of beliefs, work, and fertility*. Working Paper No. 11268. Cambridge, MA: NBER.
Fernández, R., and A. Fogli. 2006. Fertility: The role of culture and family experience. *Journal of the European Economic Association* 4: 552–561.
Fernández, R., A. Fogli, and C. Olivetti. 2004. Mothers and sons: Preference formation and female labor force dynamics. *Quarterly Journal of Economics* 119: 1249–1299.
Greenwood, J., and N. Guner. 2005. *Social change*. Economie d'Avant Garde Research Reports No. 9, Economie d'Avant Garde.
Greenwood, J., A. Seshadri, and M. Yorukoglu. 2005. Engines of liberation. *Review of Economic Studies* 72: 109–133.
Greif, A. 1994. Cultural beliefs and the organization of society: A historical and theoretical reflection on collectivist and individualist societies. *Journal of Political Economy* 102: 912–950.
Greif, A. 2005. *Institutions: Theory and history. Comparative and historical institutional analysis*. Cambridge: Cambridge University Press.
Guiso, L., P. Sapienza, and L. Zingales. 2003. People's opium? Religion and economic attitudes. *Journal of Monetary Economics* 50: 225–282.
Guiso, L., P. Sapienza, and L. Zingales. 2005. *Cultural biases in economic exchange*. Working Paper No. 11005. Cambridge, MA: NBER.
Guiso, L., P. Sapienza, and L. Zingales. 2006. Does culture affect economic outcomes? *Journal of Economic Perspectives* 20(2): 23–48.
Hanushek, E., and D. Kimko. 2000. Schooling, labor-force quality, and the growth of nations. *American Economic Review* 90: 1184–1208.
Henrich, J., R. Boyd, S. Bowles, C. Camerer, E. Fehr, H. Gintis, and R. McElreath. 2001. In search of homo economicus: Behavioral experiments in 15 small-scale societies. *American Economic Review* 91(2): 73–78.
Ichino, A., and G. Maggi. 2000. Work environment and individual background: Explaining regional shirking differentials in a large Italian firm. *Quarterly Journal of Economics* 115: 1057–1090.
Mailath, G., and A. Postlewaite. 2003. The social context of economic decisions. *Journal of the European Economic Association* 1: 354–362.
Manski, C. 2000. Economic analysis of social interactions. *Journal of Economic Perspectives* 14(3): 115–136.
Marmot, M.G., S.L. Syme, A. Kagan, H. Kato, J.B. Cohen, and J. Belsky. 1975. Epidemiologic studies of coronary heart disease and stroke in Japanese men living in Japan, Hawaii and California: Prevalence of coronary and hypertensive heart disease and associated risk factors. *American Journal of Epidemiology* 102: 514–525.
Stigler, G., and G. Becker. 1977. De gustibus non est disputandum. *American Economic Review* 67: 76–90.
Tabellini, G. 2005. *Culture and institutions: Economic development in the regions of Europe*. Working Paper No. 1492. Munich: CESifo.

Cumulative Causation

Carlos J. Ricoy

The notion of 'cumulative causation' constitutes a basic hypothesis on the workings of the market mechanism. The operation of markets is conceived as a continuous process in which economic forces interact upon one another in a cumulative way, thus making for changes in one direction to induce supporting changes which push the system further away from its initial position. In essence, this is the notion which Myrdal refers to as the 'principle of circular and cumulative causation' and which plays an organizing role in his analysis of 'uneven development' (Myrdal 1957).

Although the term 'cumulative causation' is due to Myrdal, the basic hypothesis appears in Young's analysis of 'economic progress' (Young 1928). It is on this basis that Kaldor puts forward a definite 'cumulative causation' approach to the

'economic process' (Kaldor 1966, 1967, 1970, 1972, 1974, 1975, 1978a, 1978b, 1981a, 1981b, 1985); a similar approach, even if less developed, is found in Svennilson's analysis of economic growth (Svennilson 1954).

Young's Increasing Returns

Young's increasing returns constitute the dynamic counterpart of Adam Smith's dictum 'the division of labour – cause of the increased productive powers of labour – is limited by the extent of the market' (Smith 1776). In Young's interpretation, the expansion of markets leads to an 'increasing use of roundabout methods of production' and to a 'progressive division and specialization of industries' which result in a rising 'efficiency of production' (Young 1928).

The 'progressive division and specialization of industries' refers to the tendency (implied by the expansion of markets) for industries to be broken up into more specialized concerns which concentrate on a narrower range of output.

The growth of markets makes possible a progressive 'horizontal diversification' of consumer goods industries that relates to the introduction of new products and the increasing differentiation of basically the same type of goods. As for intermediate and capital goods industries, the process of specialization is both 'horizontal' and 'vertical' ('vertical disintegration', Stigler 1951). To a large extent, this process depends on the occurrence of 'technological convergence' which itself depends on the expansion of markets. (Hirschman, 1957; Rosenberg 1976; Kaldor 1985). This notion refers to the accumulated backward linkages of industries at a given stage in the network of interindustry relations which come to share basically the same process of production, thus allowing and inducing the progressive specialization of industries at successive lower stages.

In this view, 'efficiency' appears as a '*dynamic, macroeconomic-structural*' phenomenon; for it relates to the processes of mechanization and structural transformation which, in turn, refer to the expansion of manufacturing as a whole. This, however, does not imply that efficiency is uniform across industries. On account mainly of the differential incidence of the process of mechanization and of technical progress, the '*opportunity for efficiency*' varies widely across industries which, indeed, gets reflected both in the level and in the rate of change of efficiency. Notionally, for each industry, 'opportunity' defines a 'standard' both for the level and for the rate of change of efficiency. The actual performance of the different industries can then be measured relative to the respective standards.

Relative efficiency gains in a given industry depend on mechanization and specialization which depend on the growth of markets. On account of manufacturing's internal linkages and of 'technological convergence', the growth of markets depends on the growth and specialization of other industries which, in turn, depend on the growth and specialization of yet other industries, and so on; thus, the relative rise in efficiency in an industry depends on the overall expansion of manufacturing. At the same time, the specialization of an individual industry is but the result of the general process of 'division and specialization of industries'; to a large extent, an industry gets specialized, comes to concentrate on a narrower range of output insofar as other industries do. (complementary and subsidiary industries).

On this account, therefore, an industry's (relative) rise in efficiency depends on the expansion and internal development of manufacturing as a whole as much as it depends on the internal development of the industry itself. Moreover, due to the interindustry linkages internal to manufacturing, the rise in efficiency that notionally refers to a particular industry gets spread to other industries and, eventually, to the whole sector. In this sense, the development of particular 'key' industries characterized by a high 'opportunity' for efficiency and/or by rapidly growing markets benefits the development and efficiency of other industries across the sector; in this way, 'mature' industries may benefit and receive a new 'lease of life' from developments initiated elsewhere within the sector.

Efficiency, as macroeconomic phenomenon, reflects the process of growth in terms of capital accumulation and structural transformation; in the

process, as different 'tensions, disproportions and bottlenecks' are continuously being solved, the efficiency of production rises all across manufacturing. At any one time, the level of efficiency is the reflection of the structure of manufacturing in terms of the degree of mechanization and of specialization and diversification of industries as well as in terms of the strength of the network of interindustry relations. In this sense, the level of efficiency is an index of the level of development and the rate of change is an index of the growth and development performance over a period.

On this basis, a given industry located in different countries will tend to show in each country a level and a rate of change of efficiency in correspondence with the overall development (and efficiency) and the rate of output growth (and efficiency gains) of the respective national manufacturing sectors. Thus, countries experiencing (relative to their competitors) higher rates of growth and transformation and, therefore, higher rates of efficiency as regards manufacturing as a whole will tend to experience higher corresponding rates as regards individual industries as well (see Eatwell 1982). Yet, a proviso must be made here; for 'opportunity for efficiency' varies across countries as well. The degree of 'opportunity' is related to the dependence of efficiency on learning, on the accumulation of experience and mastery and on technical progress. As regards individual industries, a high degree of opportunity results from the industries' efficiency being dependent on (intensive in) learning, mastery and technical progress. As regards countries, the opportunity for efficiency, as it depends on learning and technical progress, is related to the growth and development of manufacturing. Thus, the intercountry differential in 'opportunity' is related to the interindustry differential; the lower the opportunity of the industry is, the lower the intercountry differential and, vice versa, the higher the opportunity of the industry, the higher the intercountry differential. Thus, the correspondence between the rate of change of efficiency in manufacturing and that in individual industries will tend to be more accurate for high opportunity industries.

Say's Law as 'Closure' of the System

In Young's analysis, based on the 'classical version' of Say's Law, the growth of markets is defined by the rise in the volume of production, which, in turn, is determined by the rise in efficiency; on account of increasing returns, the latter is determined by the growth of markets itself. Hence, 'the growth of markets is determined by the growth of markets'. This, as Young points out, 'is more than mere tautology'; the expansion of markets leads to the rise in efficiency through mechanization and structural transformation which open up '*new opportunities for further change which would have not existed otherwise*'. In the normal operation of markets any given 'impulse' is amplified cumulatively, the growth of demand results in an endless 'chain reaction' of sectoral supplies and demands all through the network of interindustry relations. In the process, 'each sector receives impulses' for change and, in turn, 'sends impulses' for further change. Thus, '*change becomes progressive and propagates itself in a cumulative way*' (Young 1928; Kaldor 1972, 1973).

Young's analysis embodies the essence of the principle of cumulative causation; however, what is essentially a matter of impulses and inducements what is a 'potentiality' is transformed into 'actuality'; for, by definition, demand always responds to the inducement to further growth provided by structural change and the all-round improvement in efficiency; the actuality of an endless chain of circular and cumulative causation is thus ensured. Due to Say's Law, there is no 'degree of freedom', no independent leading element in the system. In this sense, the analysis remains as it were 'hanging in the air'.

Learning and Technical Progress

(a) To a large extent, economic growth is to be seen as a '*learning process*' (Rosenberg 1976). Economic growth involves a series of ongoing activities, decisions and events, ranging from the operation of plant and equipment and decisions taken about production and investment to the

occurrence of structural change and the introduction and development of technology. As those activities, decisions and events materialize, different problems, 'tensions, disproportions and bottlenecks', are encountered. Successful growth requires that solutions be found to those problems; and learning results from discovering and facing problems and from searching for and finding solutions to them. Moreover, the faster growth takes place, the more will problems assert themselves, the more will the need to solve them be felt and, therefore, the stronger will the inducement to learn be (see Arrow 1962). To the extent that learning is effective, experience and knowledge are accumulated and skills and capabilities developed. On this basis, each successive step in the normal process of growth and transformation becomes (potentially) easier. As a result, the efficiency of production and that of the growth and learning processes themselves are raised. Thus, learning depends on the growth of output, results in efficiency gains and induces further growth and transformation.

(b) Within the cumulative causation framework, normal technical progress is conceived of as a *non-random, evolutionary* process; it is '*non-random*' in that the direction of change is defined by the 'state of the art of the technologies already in use' and it is '*evolutionary*' in that it normally involves the 'rejection of parts of the old technology' rather than its total rejection. Moreover the process tends to be *cumulative*; for the likelihood of success in the completion of a new technology depends on past developments and on the accretion of experience, knowledge and technological mastery that results from the 'learning process' that the development of technology implies (see Dosi 1984; Rosenberg 1976; Nelson and Winter 1982).

A new technological concept (product, equipment, technique) does not come about as a 'perfectly known', 'fully grown' output of R&D activities. At this stage, there is a fundamental lack of understanding and a great deal of uncertainty as to the actual performance of the new concept. Most commonly the resolution of that uncertainty and the acquisition of knowledge require experience in the production and operation of the new concept as well as further research and development.

At the same time, the introduction of a new technological concept faces fundamental uncertainties as to the nature of demand. Consumers are both uncertain about their preferences and unaware of the characteristics of new products and of how they compare with possible substitutes. Users of capital equipment are, in turn, uncertain as to the exact nature of their requirements and lack basic knowledge and face fundamental uncertainty as to the characteristics of new technological alternatives and as to how they fit in the process of production. It is only through actual experience in consumption (Pasinetti 1982; 'product-cycle'model: Vernon 1966) and in the effective use of equipment in production that uncertainty can be resolved and knowledge increased. ('learning by using': Rosenberg 1982).

Technical progress manifests itself in a *sequence of problem-solving activities* along the chain '*demand–production–R&D*'. In this sequence, faults, weak spots and technical problems of a new technological concept will be discovered. To the extent that solutions are found and that, through operating experience, consumers/users are able to improve the specification of their preferences/requirements, there will be frequent modifications and improvements of the new concept. At the same time the discovery of problems, the search for and the actual finding of solutions will result in a better understanding of the new technology and in the accumulation of skills, experience and knowledge; in this sense, technical progress is in itself a '*learning process*'.

So technical progress depends on the dynamics of demand in a fundamental manner. On the one hand, technical progress is, to a large extent, *induced* by the expectation of demand. As regards capital goods industries there is an 'external impulse', signalled by the rate of investment, to accommodate specific requirements of different industries, particularly, of fast growing, highly innovative ones (see Schmookler 1966; Rosenberg 1976). As for consumer goods, technical change is induced by the expected evolution of consumer's 'wants' and by the expectation of extending given patterns of consumption to

lower brackets of the income distribution structure. At the same time, the success (and further development) of technological advances depends on the actual dynamics of demand; the effective development of technology requires that the expectations of demand be fulfilled, i.e. it requires the *validation* of the effective growth of markets.

On the other hand, the *efficiency* and *effectiveness* of the process of technical change depend largely on the growth of markets. A faster growth of markets makes it easier to ascertain the 'new ways of expansion' as well as to switch from one path of expansion to another; in this sense, it provides the innovative process with a '*higher degree of flexibility*' which lowers the risks and costs involved, thus leading to a higher rate of innovative effort. In addition, a faster growth of markets leads to a faster and more effective learning process; it leads to a higher rate of accumulation of skills, knowledge and mastery and, therefore, to a more efficient and effective handling of the sequence of problem-solving activities that the process of technical change entails.

The process of technical progress appears intrinsically connected to Young's increasing returns; for, both through the introduction and differentiation of goods and through the development of new technologies of production, it induces and, at the same time, is induced by the processes of mechanization and of division and specialization of industries.

Effective Demand

By combining the dynamics of 'efficiency' as it results from Young's increasing returns and from learning and technical progress with the principle of effective demand in a dynamic setting, Kaldor provides the definitive development of the principle of cumulative causation. On this account, the growth of demand constitutes the 'leading factor' of the 'self-reinforcing dynamics' internal to manufacturing. The growth of demand determines the growth of output and leads to a rising efficiency of production; whether the process keeps its momentum and becomes cumulative or gets stopped (and probably reversed) depends on the 'next round' of demand, on the response of demand to the *inducement* to further growth provided by the rise in efficiency. In this sense, the growth of demand (as 'leading factor') is the 'weak-link' of the internal dynamics of manufacturing. Thus, effective demand provides the circular process of cumulative causation with a 'degree of openness' that contrasts with the 'continuity' that, owing to Say's Law, Young's analysis implied.

In the last analysis, the key role of demand rests on the 'independence' of capital accumulation as the driving force of the process of economic growth; capital accumulation is central to technical progress and adds both to demand and to capacity, i.e. it 'provides the incentives and the means of further expansion' (Kaldor 1966). On this basis, the *potential* for a continuous self-expansion of the system finds no limit; as Joan Robinson puts it, 'carrying itself by its own bootstraps is just what a capitalist economy *can* do' (Joan Robinson 1962).

The fact that investment is the fundamental independent variable of the system does not mean that the rate of accumulation is fixed, invariant with respect to economic conditions; actually, the assumed degrees of freedom in investment behaviour are to a large extent the reflection of the many factors, economic and non-economic, that influence investment decisions. In the context of the 'self-reinforcing dynamics' of manufacturing it is the response of investment to the growth of markets and to the resulting rise in efficiency along with the more direct dependence of consumption on industrial expansion that accounts for the 'reverse link' and, thus, for the circular and cumulative nature of the dynamics itself.

The rise in (quantitative) efficiency that results from industrial expansion, leads to the growth of consumer demand through the rise in real income and through the expansion of consumption to lower brackets of the income distribution structure. Owing to the different growth elasticities of demand for different goods and for different income levels, the overall growth of consumption and, therefore, the overall growth effect of the interaction 'industrial expansion–consumption' depend largely on the composition of

consumption in terms both of goods and of income groups. In this regard, at high income levels there is a tendency for the growth of demand to slow down and, even, to stagnate; this tendency, however, is continuously being overcome by the 'innovation' and 'quality-differentiation' effects of efficiency (technical change, structure diversification); these effects are, thus, crucial to keep the 'drive' for expansion alive (Pasinetti, 1982). As regards capital goods, the 'price' and 'quality-obsolescence' effects of technical change constitute fundamental determinants of investment activity. In addition, investment is induced by industrial growth itself both through the interindustry expansion of markets and through the growth of consumption it entails.

In the sequence 'industrial growth – consumption – investment – industrial growth', the growth of real wages is of special significance; for wages are both income and cost of production. (Kalecki, 1939). As income, the growth of real wages leads to demand and output growth which lead to productivity growth; as a result, it may lead to higher rates of investment. As cost, the growth of wages has two contradictory effects. On the on hand, it may induce a process of dynamic substitution à *la* Marx, thus, leading to a higher rate of investment and productivity. On the other hand, to the extent that the growth of real wages 'eats up' in profits, it may result in a lower rate of investment. On the basis of the 'substitution' and 'demand' effects, the growth of real wages can be seen to imply a 'cumulative causation' pattern of growth. The growth of real wages leads to demand growth and to dynamic substitution, therefore to investment and to productivity growth which, in turn, lead to higher growth of wages and so on and so forth.

Manufacturing as the Engine of Growth

If capital and labour, as resources used in production, were exogeneously given in fixed quantities, the circular and cumulative dynamics of manufacturing would not materialize; economic growth would be effectively 'resource-constrained'. Yet the quantity and quality as well as the sectoral distribution of labour and capital cannot be taken as 'given' independently of the growth progress itself.

Insofar as the effective labour force (participation) varies in direct relation with demand and, most significantly, insofar as there is *surplus labour* in other sectors such as agriculture and services, at no time can the labour force be considered fully and optimally employed. As manufacturing expands, labour is drawn from those 'reserves' and gets allocated to other uses where its contribution to the economy's output is greater than before; moreover, in the process, 'efficiency' in the surplus labour sectors is enhanced. As for capital, in no significant sense can it be regarded as a 'scarce' resource; capital, as produced means of production, is output and, as such, is the result of economic activity. The quantity, quality and sectoral distribution of capital and labour, as determined by investment, learning and technical progress, are the 'effect' of the process of development as much as the 'cause' of it (see Kaldor, 1978).

In the course of its expansion, manufacturing *generates* its own resources, it mobilizes labour and produces capital; thus, the expansion of manufacturing represents a net addition to the effective use of resources and, therefore, to the overall growth of the system. The growth of manufacturing and, thus, overall economic growth, are not constrained by resources; rather, they are led by the expansion of markets, by the growth of demand for manufactures. Moreover, as other sectors depend largely on manufacturing for the provision of their inputs, the overall efficiency of the economy is mainly determined by that of manufacturing.

Economic growth, centred on the 'self-reinforcing dynamics' internal to manufacturing, is a circular process of cumulative causation governed by the growth of demand. The growth of demand for manufactures leads to the growth of output and to efficiency gains both in manufacturing and in the economy as a whole (capital accumulation-mechanization, structural transformation, reallocation of resources, learning and technical change) which *induce* further growth of demand and so on. In this view, a 'pause' in

the expansion of the system may well lead to (structural) stagnation; for, on account of dynamic economies of scale, a shortage of effective demand tends to get amplified throughout the economy. Such a cumulative downturn is but the manifestation of the free workings of the market as mechanism that *transmits impulses and inducements*; in the same way as change calls forth supporting change, the absence of change leads to stagnation; for it is in the nature of the manner of operation of markets and competition that 'growth requires growth', 'success requires success'; the market system is prone to cumulative movements, once in a growth path it tends to cumulative self-expansion; yet should growth lose its momentum and slow down, a tendency towards a cumulative downturn ensues; 'a capitalist economy cannot afford to stay still because, it it stops expanding, it falls back' (Pasinetti, 1982). This is circular and cumulative causation. (Compare Marx's account of the process of capitalist development and of competition as a dynamic process.)

Foreign Trade

At a more concrete level of analysis, as regards the 'open economy', foreign trade in manufactures asserts itself as a 'built-in' element making for the continuity of the circular process of cumulative causation. The growth and composition of net exports (exports minus imports) are fundamental elements in shaping the process of growth and structural change of the economy and, in turn, as determined by competitiveness which results from efficiency, they are mostly determined by economic growth itself. The growth and the change in the composition of demand, including those of net exports, lead to 'efficiency gains' which, via price and non-price factors, give competitiveness which leads to growth and changes in the composition of net exports and, thus, of demand and output and so on and so forth.

On the other hand, foreign trade, through the balance of payments position, may impose an 'effective constraint' on growth. But for the possibility of attracting a continuous net inflow of capital, in the long run an economy's growth rate cannot be higher than the rate of growth consistent with balance of payments equilibrium on current account. This growth rate Thirlwall (1980) refers to as 'balance of payments equilibrium growth rate'. This rate depends fundamentally on the trade balance in manufactures and, therefore, on the growth rate and on the 'normal' competitiveness of manufacturing (given the 'normal' time paths of the other components of the current account). The significance of the equilibrium rate as ceiling to the actual rate is to be understood in relation to the growth rate that can be considered as 'socially necessary' in terms of economic development (output and employment growth, structural transformation learning and technical progress) (cf.Singh's notion of an 'efficient manufacturing sector': Singh 1977). In this regard, a situation in which, due to a 'weak' competitive position, the 'equilibrium' rate is lower than the 'socially necessary' rate tends to be 'self-perpetuating', as it will result in a lower rate of efficiency gains, lower growth of net exports, and, eventually, in a lower equilibrium rate.

The dynamics of foreign trade and, thus, economic growth in a given economy depend also on economic growth and efficiency abroad. Foreign growth 'complements' domestic growth insofar as it widens the opportunities for the expansion of domestic net exports. At the same time foreign growth enlarges the supply of commodities which can 'compete' with domestic production, and, by raising foreign efficiency, lowers domestic competitiveness, thus making for a lower growth of domestic net exports (Sayers 1965; Singh 1977). In addition, the competitive process in world markets entails a fundamental 'composition effect'; fast growing countries tend to gain market shares particularly in those trades with the highest 'opportunity for efficiency' and with the highest potential for market expansion while 'weak' countries tend to be pushed out of those trades. (cf. above, 'bias' in the correspondence across countries between efficiency in manufacturing and in individual industries). In this regard, if, in the course of growth, the 'competitive' aspects of growth elsewhere come to dominate, the ensuing disequilibrium may develop into a cumulative

downturn in which a lower growth of net exports leads to lower growth of output, lower rates of investment, lower rates of transformation and technical change, thus leading to a still lower growth of net exports and so on; eventually, the country will find itself 'balance of payments constrained' as the equilibrium rate falls below the 'socially necessary' rate. This is a '*vicious circle of cumulative causation*'. The cumulative fall in competitiveness and the resulting 'vicious circle' are worsened as other countries will be experiencing a '*virtuous circle*' in which high rates of growth of demand result in high rates of investment, transformation and technical progress, high growth of net exports (increasing market shares) and so on. A fundamental feature of this process rests in the fact that the 'vicious circle' results from developments initiated elsewhere, i.e. by the rise in growth rates and in efficiency elsewhere in the world system; thus, the need to expand, the need to keep the pace in the process of structural transformation and technical change appears more stringent in the presence of foreign trade. On the foregoing account, the dynamics of the world economy, as regulated by the free operation of markets, appears characterized by an inherent tendency towards *unequal growth* across countries; for, in a dynamic world where the expansion of manufacturing is characterized by the operation of 'dynamic economies of scale', where technology is 'developed' and not universally accessible, where technical progress unfolds as an evolutionary and cumulative learning process and where the dynamics of effective demand determine the dynamics of the system, any given 'competitive advantage' in terms of growth, transformation and technical change tends to be compounded through the interdependence that exists among countries via foreign trade.

The tendency towards 'unequal growth' entails a '*deflationary bias*' and an intrinsic and progressive instability in the dynamics of the world economy. As the growth of 'weak' countries is held back, the overall expansion of the world economy tends to slow down. This is normally offset by the faster growth of 'strong' countries which actually enables weak countries to grow faster than would otherwise be the case. In this sense, the long-rung deflationary tendency embodied in the dynamics of the system remains as it were 'disguised' in the very process of expansion. Yet, the pattern of uneven growth makes for the cross-country 'differential' in competitiveness to grow wider over time; in this respect, the dynamics of the world economy appears inherently and progressively unstable. Over time, a tendency develops for weak countries to experience a 'fundamental disequilibrium' (constraint) in the balance of payments. In these circumstances, an exogenous 'deflationary shock' or, simply, a slowdown in the expansion of the 'leaders', would make the constraint binding. But, even in the absence of any such shock, as long as the weak countries' loss of competitiveness is left 'to look after itself' and actually becomes cumulative, the tendency towards 'fundamental disequilibrium' will eventually materialize. In the event, a general slowdown in the expansion of markets could only be avoided by enabling weak, 'structural' deficit countries to grow at rates higher than the respective balance of payments equilibrium rate, thus effectively allowing them to run continuous current account deficits. If, on the contrary, the burden of adjustment were brought to bear upon weak countries by forcing them 'to live within their means', a generalized contraction of effective demand would ensue; the lower growth and weak countries and, thus, the lower growth of effective demand, would be spread from country to country through the operation of the dynamic foreign trade multiplier. In this way, a generalized deflationary process would set in. The fundamental point to stress here is that the occurrence of such a process constitutes the long-run consequence of the progressively increasing differential of competitiveness implied by the pattern of unequal growth, transformation and technical progress inherent in the normal dynamics of the world economy as regulated by the free operation of markets.

At the same time, the normal dynamics of the world economy are characterized by a pattern of

uneven development. In this regard, the sectoral structure of the economy is of the utmost significance. The dynamics of the different countries' net exports, as determined by relative efficiencies, reflects the process of development. At any one time, the structure of net exports reflects the past process of development in terms of Young's increasing returns, learning, technical change and both the potential for market expansion and the opportunity for efficiency characterizing the existing economic structure. In turn, the process of development and, thus, the economic structure reflect the dynamics of foreign trade.

As regards a 'representative' underdeveloped country, the economy is characterized by a low degree of mechanization, by a lack of inter- and intra-industry diversification and specialization and by a low degree of sectoral interdependence; in this sense, underdevelopment can be described in terms of the market-induced structural inability to realize Young's increasing returns. In addition, production processes are characterized by a very low intensity in skills, knowledge and technical progress which result in the absence of induced learning. On this basis, the export composition appears basically centred on primary commodities and a few 'mature' manufactured products for which the growth of demand tends to be rather low and unstable while the 'opportunity for efficiency' is practically nil.

As for a 'representative' advanced country, its economic structure, centred on the manufacturing sector, is highly diversified and interdependent; in turn, the different industries are highly specialized in terms of products and methods of production; the latter are further characterized by a high average degree of mechanization and by a high average intensity in accumulated experience, knowledge and technological mastery.

In the normal (free) operation of markets the pattern of 'specialization' as between advanced and 'poor' countries tends to be perpetuated and the corresponding differential and efficiency tends to be grow wider over time; due to the operation of dynamic economies of scale, industrial activities tend to concentrate in a few 'established centres' which benefit from the 'freeing' and 'widening' of markets at the expense of the industrial development of 'backward' countries; 'the free play of market forces works towards inequality' (Myrdal 1957).

See Also

- ▶ Increasing Returns to Scale
- ▶ Kaldor, Nicholas (1908–1986)

Bibliography

Arrow, K.J. 1962. The economic implications of learning by doing. *Review of Economic Studies* 28(3): 155–173.

Cornwall, J. 1977. *Modern capitalism: Its growth and transformation*. Oxford: Martin Robertson.

Cripps, F., and R. Tarling. 1973. *Growing in advanced capitalist economies, 1950–1970*. Cambridge: Cambridge University Press.

Dosi, G. 1984. *Technical change and industrial transformation*. London: Macmillan.

Eatwell, J. 1982. *Whatever happened to Britain?* London: Duckworth..

Hirchman, A. 1958. *The strategy of economic development*. New Haven: Yale University Press.

Kaldor, N. 1966. *Causes of the slow rate of economic growth of the United Kingdom*. Cambridge: Cambridge University Press.

———. 1967. *Strategic factors in economic development*. Ithaca: Cornell University Press.

———. 1970. The case for regional policies. *Scottish Journal of Political Economy* 17(3), November, 337–348.

———. 1972. The irrelevance of equilibrium economics. *Economic Journal* 82: 1237–1255.

———. 1974. Teoria del equilibrio y teoria del crecimiento. *Cuadernos de Economica*. (Reprinted in translation in *Economics and human welfare: Essays in honour of T. Scitovsky*, ed. M.J. Boskin, London: Academic Press. 1979).

———. 1975. Economic growth and the Verdoorn Law: comment on Mr. Rowthorn's article. *Economic Journal* 85, December, 891–6.

———. 1978a. *Further essays on economic theory*. London: Duckworth.

———. 1978b. *Further essays on applied economics*. London: Duckworth.

———. 1981a. Discussion. In *Macroeconomic analysis*, ed. D. Currie, R. Nobay, and D. Peel. London: Croom Helm.

———. 1981b. The role of increasing returns, technical progress and cumulative causation in the theory of

international trade and economic growth. *Economie Appliquée* 34(4): 593–617.
———. 1985. *Economics without equilibrium*. Cardiff: University College Cardiff Press.
Kalecki, M. 1966. *Studies in the theory of business cycles 1933–1939*. Oxford: Basil Blackwell.
Myrdal, G. 1957. *Economic theory and underdeveloped regions*. London: Duckworth.
Nelson, R., and S. Winter. 1982. *An evolutionary theory of economic change*. Cambridge, Mass.: Harvard University Press.
Nurkse, R. 1953. *Problems of capital formation in underdeveloped countries*. Oxford: Oxford University Press.
Pasinetti, L. 1981. *Structural change and economic growth*. Cambridge: Cambridge University Press.
Prebisch, R. 1950. *The economic development of Latin America and its principal problems*. New York: United Nations.
———. 1959. Commercial policy in underdeveloped countries. *American Economic Review, Papers and Proceedings* 49, May, 251–273.
Robinson, J. 1962. *Essays in the theory of economic growth*. London: Macmillan.
———. 1965. The new mercantilism. In *Collected economic papers of Joan Robinson*. Oxford: Basil Blackwell.
Rosenberg, N. 1976. *Perspectives on technology*. Cambridge: Cambridge University Press.
———. 1982. *Inside the black box: Technology and economics*. Cambridge: Cambridge University Press.
Sayers, R.S. 1965. *The vicissitudes of an export economy: Britain since 1880*. Sydney: University of Sidney.
Schmookler, J. 1966. *Invention and economic growth*. Cambridge, MA.: Harvard, University Press.
Singh, A. 1977. U.K. industry and the world economy: a case of de-industrialisation? *Cambridge Journal of Economics* 1(2), June, 113–36.
———. 1978. North Sea oil and the reconstruction of the UK industry. In *De-Industrialisation*, ed. F. Blackby. London: Heinemann.
Smith, A. 1776. In *An inquiry into the nature and causes of the wealth of nations*, ed. E. Cannan. London: Methuen 1904.
Stigler, G. 1951. The division of labour is limited by the extent of the market. *Journal of Political Economy* 59, June, 185–93.
Svennilson, I. 1954. *Growth and stagnation in the European economy*. Geneva: UN Economic Commission for Europe.
Symposium. 1983. Kaldor's Laws. *Journal of Post-Keynesian Economics* 5(3): 341–429.
Thirlwall, A.P. 1980. *Balance of payments theory and the United Kingdom experience*. London: Macmillan.
Vernon, R. 1966. International investment and international trade in the product cycle. *Quarterly Journal of Economics* 80, May, 190–207.
Young, A. 1928. Increasing returns and economic progress. *Economic Journal* 38: 527–542.

Cumulative Processes

Björn Hansson

The first well-known analysis of cumulative processes was developed by Knut Wicksell in his book *Interest and Prices*, which was published in 1898. It grew out of an attempt to reformulate the quantity theory of money. In this context the cumulative process is intimately connected with the development of the saving–investment approach, which is one mechanism through which a change in the quantity of money can influence prices and quantities.

Wicksell considered the main proposition of the quantity theory, namely, that the value of the purchasing power of money varies in inverse proportion to its quantity, to be basically correct. At the same time the quantity theory was in its original formulation too restrictive and in conflict with reality, since it was based on the assumption that everybody uses their own cash, which is both legal tender and the monetary base, for buying and selling and all have to maintain a cash balance. Therefore the followers of the quantity theory sometimes argued 'as though the quantity of money, or of that part that at any moment finds itself in the hands of the public, must act as a *direct* and *proximate* price-determining force' (Wicksell 1898, p. 43). This is the so-called direct mechanism, which works via a real-balance effect and it is still stressed by Friedman in his development of modern monetarism. However, in a developed credit economy the keeping of individual cash-balances has almost faded away and it has been replaced by current and deposit accounts and the use of claims of various kinds in monetary transactions. The banks, in their turn, only hold a smaller part of the deposited sums as cash, which shows that Wicksell developed the quantity theory for a banking system based on fractional reserves.

The Background to the Cumulative Process

The following quotation gives the analytical basis for the cumulative process:

> Every rise or fall in the price of a particular commodity presupposes a disturbance of the equilibrium between the supply and the demand for that commodity, whether the disturbance has actually taken place or is merely prospective. What is true in this respect of each commodity separately must doubtless be true of all commodities collectively. A general rise in prices is therefore only conceivable on the supposition that the general demand has for some reason become, or is expected to become, greater that the supply. This may sound paradoxical because we have accustomed ourselves, with J.B. Say, to regard goods themselves as reciprocally constituting and limiting the demand for each other. And indeed ultimately they do so; here, however, we are concerned with precisely what occurs, in the first place, with the middle link in the final exchange of one good against another, which is formed by the demand of money for goods and the supply of goods for money (Wicksell 1935, pp. 159–60).

Wicksell proceeded from an approach based on Marshallian partial equilibrium to an aggregate approach for analysing secular changes in the general price level. Furthermore, the central problem is the analysis of a system out of equilibrium (i.e. a disequilibrium analysis), which implies criticism of the quantity theory for analysing and comparing equilibrium situations only and leaving out the dynamic process itself.

The cumulative process takes its point of departure in an analysis of the relation between the actual rate of interest on loans, the money rate, and the natural rate or the normal rate. The *natural rate* is defined as the anticipated profit to be made by the use of a bank loan. The *normal rate* is defined as that particular level of the money rate which guarantees the equality of the demand for loan capital from investors and the supply of savings or loan capital from lenders; this rate corresponds to the natural rate.

The level of the *money rate* of interest is ultimately determined by the normal rate. But in the first instance, the money rate of interest is a separate variable which is set autonomously by the banks in the market for borrowing and lending of money. In this institutional setting, where bankers fix the money rate, the money supply is assumed to be endogenous and determined by demand. Therefore disturbances emanate from variations in the demand for money due to changes in the natural rate, which is then passively supported by an expansion of the money supply from the private banks, and it is not active changes in the monetary base initiated by the Central Bank which are the source of the disturbance. The equality of the money rate with the normal rate, which, according to Wicksell's analysis, implies stability of the price level and that saving equals investment, is thus a separate equilibrium condition, which was later called monetary equilibrium.

The Mechanism of the Cumulative Process

In the following example it is supposed that the natural rate increases, which may be due to an increase in productivity, while the money rate as fixed by the banks stays the same. It implies that the value of the total product at the end of the production period – the duration of the period is the same in all lines of production – has increased, but the entrepreneurs have to pay back less to the bank. They now have a surplus profit, which eventually will lead to an attempt to expand production. An expansion in real quantities is ruled out since full employment is assumed (Wicksell's analysis was concerned with changes in the price level and not with changes in quantities.) In any case, even if capital accumulates, it would still take some time before the results accrue, so there would be no immediate counteracting effects. However, the entrepreneurs will try to increase their demands for inputs and these prices will rise, and it is assumed that the increase is equal to the value of the expected surplus profit. The owners of inputs, in their turn, will increase their demand for consumer goods and the price level will rise proportionately. At the end of the year, the entrepreneurs have in their hands, at the new

price level, consumer goods of a higher value than at the end of the previous period, while their debts to the banks are the same. Hence, they will still have a surplus profit, and the tendency to expand output will persist despite the change of the price level. The whole process will repeat itself. This is the minimum requirement for denoting a process as cumulative: it should have an endogenous mechanism which keeps the process going.

If it is assumed that the entrepreneurs always anticipate that the price level for consumer goods will be the same at the beginning as well as the end of the period, then we can imagine that a steady and more or less uniform rise in the price level will ensue. In the case of entrepreneurs expecting future rises in prices, then the actual rise will be higher and the cumulative process will be more and more rapid.

In the reverse case, where the money rate is larger than the natural rate, the process will be similar since Wicksell assumed that prices and wages have the same flexibility in both directions. It is likely that full employment could be preserved during the downward process; the fall in entrepreneurial demand for labour and land will induce workers and landlords to reduce their claims for wages and rents and the activity will be maintained at its former level. Wicksell did not rule out the existence of unemployment during a downward cumulative process, but it would not be a cumulative change.

The cumulative process may come to an end through internal causes. The changes in the price level will act as an equilibrating mechanism via its effect on the level of bank reserves. The increase in prices will lead to a higher requirement of means of exchange, which implies that to maintain a rate of interest permanently below the natural rate it is necessary to increase continually the amount of reserves. The Central Bank can therefore play a role by changing the monetary base.

In the new equilibrium the current price level will not change since the entrepreneurs can pay the increased wages etc. and still earn the normal profit. This is the basis for Wicksell's discussion of the difference between a stable equilibrium of relative prices and a *neutral/indifferent equilibrium* for the general price level, which implies that there is no tendency to resume the old equilibrium position of the price level. In fact, in an equilibrium situation value theory will determine relative prices while the quantity theory of money gives the absolute height of the price level. Wicksell did not challenge the classical dichotomy as long as it is a comparison of equilibrium situations. However, during a cumulative process, which is a disequilibrium phenomenon, the connection between value theory and monetary theory proceeds via the difference between the money rate and the normal rate. This difference determines changes in the price level and not the level itself. Thus the cumulative process is in the first instance an analysis of changes in the general price level and its main concern is not with changes in output and employment. This is explicit in the formal analysis of *Interest and Prices*, where it is assumed that the real system is in a stationary equilibrium even during the cumulative process.

The Cumulative Process and Monetary Policy

Wicksell applied his reconstruction of the quantity theory to Tooke's criticism of the quantity theory; rising prices mainly coincide with rising or high interest rates, which Keynes called Gibson's paradox. The object of this debate is the secular increase in the price level, which may be exemplified by the period 1850–1873. Wicksell explained this fact by the tardiness of the banks to change the loan rate in relation to changes in the natural rate. It is thus changes in the natural rate which often trigger off the cumulative process. It has already been seen that Wicksell did not put the blame on the Central Bank and the private banks are now exempted from initiating disturbances, which are mainly due to changes in real factors affecting the private demand for money. This analysis obviously has a place for an active monetary policy where the Central Bank tries to influence the money rate so as to dampen the cumulative process.

Further Development of Cumulative Processes

It is of fundamental importance for later developments that Wicksell analysed movements in the general price level via the effects of changes in the rate of interest (both money and normal) on savings and investment, that is, the indirect mechanism.

The Wicksellian influence on Keynes's *Treatise on Money* came probably not through Wicksell's works but via Cassel's *The Theory of Social Economy*, which was translated in 1923. Keynes used the idea of a difference between the natural and the market rate or between savings and investment, and its influence on profit, which is defined as windfall gains, as the basis for his attempt to find the dynamical laws of the disequilibrium process. Like Wicksell, he criticized the long run equilibrium character of the propositions of the quantity theory, since it did not distinguish the factors which operated during the transition process between two equilibrium positions. Keynes's theoretical constructions, the Fundamental Equations, were not supposed to be a complete substitute for the quantity theory, but to add an analysis of short period situations which was most important, from a practical point of view, for studying monetary phenomena. Keynes extensively studied, in particular in the Fundamental Equations, the dynamics of the price level, but he went beyond Wicksell by analysing credit cycles, which involve the mechanics of the wage-price-employment structure. However, the incorporation of quantity changes in the disequilibrium process does not imply that Keynes had determined an equilibrium level of output and employment which is different from full employment.

Mises stressed that the difference between the two interest rates had differential impacts on the prices of consumption goods and the prices of capital goods, that is, an analysis of relative prices was incorporated in the dynamic process. Hayek started from Mises's contribution. He was not interested in the dynamics of the price level as such, because this magnitude, which is a statistical artifact, has no influence on the decisions of individuals. It is the direct influence on relative prices from changes in the money supply and the money rate which is important. These changes determine the level and the direction of production. This is a clear rejection of the proposition that changes in the money supply can only lead to disturbances via the general price level. Hayek's position is linked to the substitution of the old notion of a long run equilibrium by intertemporal equilibrium. The central problem is therefore not changes in the price level but disturbances of the equilibrium relation between the rates of intertemporal exchange. This spills over into the problem of neutral money: to define the conditions under which changes in the money supply might leave intertemporal price relationships unchanged, which is one development of the notion of monetary equilibrium.

Wicksell's Swedish followers, Lindahl and Myrdal in particular, used the notions of ex ante and ex post to determine the factors which constitute the ex post equality between saving and investment during a cumulative process. This analysis puts the difference between ex ante saving and ex ante investment as the main condition for monetary equilibrium, where the latter is a criterion of a cumulative process. They also made a thorough analysis of the definition of the normal rate which would guarantee monetary equilibrium, in an attempt to make this rate practically useful in a monetary analysis. In this analysis they used new equilibrium concepts like temporary equilibrium and sequence analysis, which became one of the hallmarks of the Stockholm School.

During the 1930s two different strands of thought superseded the cumulative process as an approach to the relation between savings and investments. On the one hand there is the macrodynamic analysis developed by Frisch and Tinbergen, which is centred on cyclical behaviour in prices and quantities and on the existence and stability of equilibrium in a dynamic system. On the other hand Keynes's *General Theory* focused on the determination of a short run equilibrium of output and employment. Comparative statics was used to study the effects of changes in

exogenous factors and Keynes showed very little interest in the accommodation process outside equilibrium.

In the wake of the neo-Walrasian interpretations of Keynes (e.g. Clower and Leijonhufvud), which imply a disequilibrium approach, there have been attempts to reinterpret and develop Wicksell's cumulative process from this point of view (e.g. Laidler 1972). The neo-Walrasians stress income-constrained aggregate demand functions and a dynamic process with trading at disequilibrium prices, which is opposed to the traditional tâtonnement process with an auctioneer and recontracting. From this angle, Wicksell's modern element lies in the explicit analysis of the behaviour of an economy in disequilibrium and with no recontracting, which generates an income constrained process. However, the assumption of flexible wages and prices, in particular downwards, leads only to a cumulative process in the price level while output is constant. According to Laidler (1972), it is enough to introduce the Keynesian assumption of rigid wages and sticky prices into Wicksell's cumulative process and a Keynesian income constrained process would follow immediately. In this process changes in quantities do all the adjusting necessary in disequilibrium, which may produce cumulative forces that tend to move employment away from full employment equilibrium.

See Also

- ▶ Wicksell, Johan Gustav Knut (1851–1926)
- ▶ Wicksell's Theory of Capital

Bibliography

Haberler, G. 1937. *Prosperity and depression*, 5th ed, , 1964. London: George Allen & Unwin.

von Hayek, F. 1931. *Prices and production*. London: George Routledge & Sons.

Keynes, J.M. 1930. *A treatise on money*. vols I–I-I. Reprinted in vols V–VI of *the collected writings of John Maynard Keynes*. London: Macmillan, 1971.

Laidler, D. 1972. On Wicksell's theory of price level dynamics. *Manchester School of Economics and Social Studies* 40(2): 125–44.

Lindahl, E. 1930. *Penningpolitikens medel*. Lund: C.W.K. Gleerup. Trans. as The rate of interest and the price level in Lindahl, *Studies in the theory of money and capital*. London: George Allen & Unwin, 1939.

Milgate, M. 1979. On the origin of the notion of 'intertemporal equilibrium'. *Economica* 46: 1–10.

von Mises, L. 1912. *Theorie des Geldes und der Umalaufsmittel*. Munich: Duncker & Humblot; 2nd edn, 1924. Trans. as *The theory of money and credit*. London: Jonathan Cape, 1934.

Myrdal, G. 1931. Om penningteoretisk jämvikt. En studie över den 'normala räntan' i Wicksells penninglära. *Ekonomisk Tidskrift* 33. Trans. as *Monetary equilibrium*. London: Hodge, 1939.

Patinkin, D. 1952. Wicksell's 'cumulative process'. *Economic Journal* 62: 835–47.

Patinkin, D. 1956. *Money, interest, and prices*, 2nd ed. New York: Harper & Row, 1965.

Uhr, C.G. 1960. *Economic doctrines of Knut Wicksell*. Berkeley/Los Angeles: University of California Press.

Wicksell, K. 1898. Penningeräntans inflytande på varuprisen. *Nationalekonomiska föreningens förhandlingar*, vol. I. Trans. as The influence of the rate of interest on commodity prices, in Wicksell, *Selected Papers on Economic Theory*. London: George Allen & Unwin, 1958.

Wicksell, K. 1898. *Geldzins und Gütepreise*. Jena: Gustav Fischer. Trans. as *Interest and prices*. London: Macmillan, 1936.

Wicksell, K. 1919. Professor Cassels nationalekonomiska system. *Ekonomisk Tidskrift* 21. Trans. as Professor Cassel's system of economics, in Wicksell (1934–1935).

Wicksell, K. 1934–5. *Lectures on political economy*. Vols I–II. Trans. from the third Swedish edition of *Föreläsningar i nationalekonomi*, vols I–II, 1928–1929. London: George Routledge & Sons.

Cunningham, William (1849–1919)

O. Kurer

Keywords

Cunningham, W.; Economic history; English historical school; Imperialism; Marshall, A.; Protectionism

JEL Classifications

B31

A member of the English Historical School, Cunningham was educated at the Universities of Edinburgh and Cambridge. He held various posts as lecturer at Cambridge and was elected Fellow of Trinity College in 1891. From 1891 to 1897 he was Tooke Professor of Statistics of King's College London. In addition, he pursued a religious career. He was ordained in 1874 and rose to be Archdeacon of Ely (1907–19).

Cunningham was one of the most important pioneers in economic history. His *Growth of English Industry and Commerce* (1882) was the first textbook in the field, widely used for several decades and an important foundation on which English economic history was to be constructed, and he relentlessly fought for the recognition and establishment of economic history as an independent discipline.

Cunningham became increasingly hostile towards economic theory. He felt that its assumptions about human behaviour and the institutional framework were leading to insufficiently complete analyses and were blatantly unrealistic for most periods in history. In 1892 he started the English Methodenstreit by attacking Marshall for constructing economic history from general principles instead of empirical data. The debate was partly the result of his personal and professional antagonism towards Marshall and his wish to apply economics to politics.

Cunningham shifted from an internationalist and free trader to a nationalist and protectionist, making the preservation and strengthening of the nation-state his most weighty political and economic objective. By the time of the fiscal controversy in 1903 he fully endorsed the tariff reform movement and subscribed to imperialism, with the great empire securing peace and order.

Selected Works

1882. *The growth of English industry and commerce.* Cambridge: Cambridge University Press.

1892a. The perversion of economic history. *Economic Journal* 2: 491–506.

1892b. The perversion of economic history. A reply to Professor Marshall. *Pall Mall Gazette,* 29 September, and *Academy,* 1 October, 288.

1904. *The rise and decline of the free trade movement*. Cambridge: Cambridge University Press.

1911. *The case against free trade.* Preface by Joseph Chamberlain. London: John Murray.

Bibliography

Cunningham, A. 1950. *William Cunningham: Teacher and priest*. Preface by F.R. Salter. London: Society for Promoting Christian Knowledge.

Foxwell, H.S. 1919. Archdeacon Cunningham (obituary). *Economic Journal* 29: 382–390.

Maloney, J. 1976. Marshall, Cunningham, and the emerging economics profession. *The Economic History Review* 29: 440–451.

Maloney, J. 1985. *Marshall, orthodoxy and the professionalisation of economics*. Cambridge: Cambridge University Press.

Scott, W.R. 1920. William Cunningham, 1849–1919. *British Academy, Proceedings* 9: 465–474.

Semmel, B. 1960. *Imperialism and social reform: English imperial thought 1895–1914*. London: George Allen & Unwin.

Wood, J.C. 1983. *British economists and the Empire, 1860–1914*. Beckenham: Croom Helm.

Cunynghame, Henry Hardinge (1848–1935)

John K. Whitaker

Soldier, lawyer, civil servant, polymath and amateur economist, Sir Henry Cunynghame was born of distinguished forebears on 8 July 1848 at Penshurst. He died at Eastbourne on 3 May 1935, having been knighted in 1908. In 1870 he entered St John's College, Cambridge, to study law, throwing over a promising military career. There he became a favourite of Alfred Marshall and was infected by an enthusiasm for 'geometrical political economy', a topic on which he was eventually to publish one of his many books (1904). There too he invented for Marshall a

machine (now lost) for drawing a grid of rectangular hyperbolae (Guillebaud 1961, Vol. II, pp. 37–8).

Called to the Bar in 1875, Cunynghame had a varied career in law and government, but always retained his interest in economics. He occasionally lectured on the subject (his *Notes on Exchange Value* (1880) were printed for one such course) and in the later 1880s belonged to the economic discussion group which met at the Hampstead home of Henry Ramée Beeton. (P.H. Wicksteed, G.B. Shaw, H.S. Foxwell and F.Y. Edgeworth were among the regulars.) There he presented a paper (1888) defending Marshall's supply curve against Wicksteed's criticisms. The analysis of external effects in production and consumption, his most significant theoretical contribution, first appeared here, the arguments being amplified, but not much clarified, in Cunynghame (1892). His other notable contribution, the use of back-to-back demand–supply diagrams to analyse markets linked by trade, appeared in Cunynghame (1903). The 1904 book, although lively and praised by J.M. Keynes, added little and, indeed, rather compounded earlier ambiguities by a certain flabbiness of thought. Cunynghame's last economic publication (1912) was a valedictory address on methodology. For further biographical detail see Keynes (1935) and Ward and Spencer (1938). Consult also letters by Marshall (reproduced in Pigou 1925, pp. 447–452; Guillebaud 1961, Vol. II, pp. 809–813) and Edgeworth's review (1905) of Cunynghame (1904).

Selected Works

1880 *Notes on exchange value.* London, privately printed.

1888. *Some remarks on demand and supply curves, and their interpretation.* London, privately printed.

1892. Some improvements in simple geometrical methods of treating exchange value, monopoly, and rent. *Economic Journal* 2: 35–52.

1903. The effect of export and import duties on price and production examined by the graphic method. *Economic Journal* 13: 313–323.

1904. *Geometrical political economy.* Oxford: Clarendon Press.

1912. Address to the economic science and statistical section of the British association for the advancement of science. *Journal of the Royal Statistical Society* 76: 88–98.

References

Edgeworth, F.Y. 1905. Review of Cunynghame (1904). *The Economic Journal* 15: 62–71. Reprinted in F.Y. Edgeworth, *Papers Relating to Political Economy*, vol. 3. London: Macmillan, 1925.

Guillebaud, C.W, ed. 1961. Alfred Marshall, *Principles of economics* Ninth (Variorum) Edition. London: Macmillan.

Keynes, J.M. 1935. Obituary: Sir Henry Cunynghame. *The Economic Journal* 45(178): 398–406. Reproduced in J.M. Keynes, *Collected writings*, vol. X (Essays in Biography). London: Macmillan.

Pigou, A.C. (ed.). 1925. *Memorials of Alfred Marshall*. London: Macmillan.

Ward, C.H.D., and C.B. Spencer. 1938. *The unconventional civil servant: Sir Henry Cunynghame*. London: Michael Joseph.

Currencies

C. A Gregory

Any commodity is capable of being used as a medium of exchange and history furnishes us with an almost endless list of commodities that have been used in this way: cattle, cacao, beans, salt, silk, furs, tobacco, dried fish, wheat, rice, olive oil, cloth, cowry shells, iron, copper, silver and gold. However, not all of these commodities are efficient means of exchange and the natural properties of cowry shells, gold, silver and copper led to their emergence as the most popular form of currency. The relative scarcity of gold and its indestructibility made this form of money highly desirable as a store of value; the relative abundance of cowry shells, combined with their uniformity and divisibility, made them an ideal medium of exchange for low valued transactions;

silver and copper have natural properties that made them useful for transactions intermediate between these two extremes.

The rise of nation states and the emergence of paper money and metallic tokens led to a decline in the use of commodity monies as currency. The principal problem with gold, silver, copper and cowries was that even though their relative values remained remarkably stable for centuries these ratios were incapable of being fixed. For example, the exchange ratio of silver to gold in Europe deviated little from its average of 13 to 1 in the eight hundred years prior to the emergence of the gold standard towards the end of the 19th century (Shaw 1895); in West Africa the cowry/gold ratio varied between 15,000 and 20,000 to 1 over the period 1700 to 1850 (Johnson 1970, p. 334). The ratios were determined by the conditions of production, unlike the relation of the English penny to the pound which was fixed by government fiat at 240 to 1.

Fixed ratios of the various units of a currency are essential if it is to perform its function as a standard of price efficiently. Thus 'bad' state money, which has no intrinsic value but fixed ratios between its subordinate parts, forced out the 'good' commodity monies according to the principles of Gresham's Law. Cowry shells were first to go and they disappeared from Africa and Asia in the mid to late 19th century following the colonization of the regions by the British empire and the establishment of the Sterling currency area. Silver followed next but only after much resistance from the European countries whose currencies were based on the silver or bimetal standard. Gold, too, disappeared from circulation but only to take up residence underground where central banks clung to it as a store of value. It too seemed to be on the way out in the post World War II period as the US dollar emerged as world currency. However, the rapid rise in the US dollar price of gold following the deregulation of the gold market in 1971 has effectively remonetized gold. In 1982, for example, gold accounted for only 9% of total foreign reserves held by all countries. However, this estimate values gold at the official rate of 35 SDR. If market price values are used (375 SDR) the percentage of gold to total reserves jumps to 51 per cent. Gold will no doubt continue to fulfil this function until a world government establishes a world paper currency. As there is no prospect for this happening for some time yet, gold is likely to be around for some time as a symbol of the anarchy and mistrust that characterizes the world economy.

The rise to dominance of paper money and coin tokens at the expense of metallic currencies is not difficult to understand. It was a technological advance in the sphere of exchange that was brought about by the technological changes in the sphere of production: the upsurge in the production of commodities required the development of an efficient medium for their exchange and distribution. What is difficult to explain, however, is why some 'archaic' currencies have not only continued to exist but have flourished under the impact of colonization. For example, the establishment of one of the world's largest copper mines on Bougainville Island, Papua New Guinea, has generated an enormous upsurge in demand for traditional shell currency. This has to be imported from Malaita Islands in the neighbouring Solomon Islands, some 550 kms away (Connell 1977). This is not an isolated case as similar evidence comes from other parts of the country (e.g. Chowning 1978).

At one level this can be seen as a minor problem of trying to understand the process by which a very small country of some three million people has been integrated into the world economy. At another level, however, it challenges us to reflect on the nature of our own money and society and to examine it within a comparative context. Early attempts to come to terms with the problem did so within the framework of a 'primitive money/modern money' dichotomy (Einzig 1948; Quiggin 1949). This formulation, which still has its adherents today (e.g. Melitz 1974), confuses the issue by labelling as 'primitive' what is obviously very much a 'modern' phenomenon. What is needed is a theoretical framework that gets beyond the terminological and conceptual inadequacies of the 'primitive/modern' dichotomy. Karl Polanyi and his followers have made important contributions in this regard (see Polanyi 1977). The 'gift/commodity' distinction (Gregory 1982), which has its origins in the

theories of Mauss (1925) and Marx (1867) respectively, is a recent attempt to come to terms with the problem from a somewhat different theoretical perspective. The issues here are complex and controversial and the debates surrounding this issue will no doubt continue for some time to come.

See Also

- ▶ Commodity Money
- ▶ Fiat Money
- ▶ Fiduciary Issue
- ▶ Money Supply

Bibliography

Chowning, A. 1978. Changes in west New Britain trading systems. *Mankind* 11: 296–307.
Connell, J. 1977. The Bougainville connection: Changes in the economic context of shell money production in Malaita. *Oceania* 48(2): 81–101.
Einzig, P. 1948. *Primitive money*. London: Eyre & Spottiswoode.
Gregory, C.A. 1982. *Gifts and commodities*. London: Academic Press.
Johnson, M. 1970. The cowrie currencies of West Africa. *Journal of African History* 11(17–49): 331–353.
Marx, K. 1867. *Capital*. Moscow: Progress.
Mauss, M. 1925. *The gift*. London: Routledge.
Melitz, J. 1974. *Primitive and modern money*. Reading: Addison-Wesley.
Polanyi, K. 1977. *The livelihood of man*. London: Academic Press.
Quiggin, A.H. 1949. *A survey of primitive money*. London: Methuen.
Shaw, W.A. 1895. *The history of currency. 1252–1894*. London: Wilson & Milne.

Currency Boards

Federico Sturzenegger

Abstract

Currency boards are exchange rate arrangements in which the exchange rate is fixed to an anchor currency and central banks just buy and sell domestic currency at this exchange rate. We review the advantages and disadvantages of currency boards. While some of the alleged benefits of currency boards have diminished hand in hand with a reduction in inflation rates in most countries since the mid-1990s, currency boards may remain an attractive option for certain countries.

Keywords

Bank crises; Central bank independence; Commitment; Crredibility; Currency boards; Currency unions; Dollarization; Exchange rate policy; Foreign exchange markets; Inflation; Inflation targeting; Inflation expectations; International reserves; Lender of last resort; Monetary base; Monetary policy; Money supply; Optimal currency area; Seigniorage

JEL Classifications

F3

A currency board is defined as an exchange rate arrangement in which the exchange rate is fixed to an anchor currency and the central bank operates with a simple rule that precludes the monetary authorities from issuing money unless they obtain an equivalent amount of international assets to back it. From a practical point of view this means that the central bank has no independent monetary policy and that it creates or contracts the money supply only as the result of its interventions in the foreign exchange market. If there is excess demand for domestic currency capital will flow in (probably in response to an increase in interest rates) and the central bank, by acquiring these flows, will expand the money supply. If there is excess supply of domestic currency, the central bank will take in this excess supply by giving away international assets, thus contracting the money supply. In some cases this rule is implemented by forcing the central bank to have full backing of domestic base money with international reserves. In some cases a currency board does not require a one-to-one backing of the monetary base, but it still precludes the conduct on an independent monetary policy beyond very strict

limits. In fact, a currency board also differs from a typical peg in its commitment to the system, which is usually enshrined in law and in the Central Bank charter.

As of July 2006 the exchange rate arrangement classification published by the International Monetary Fund (IMF) identifies 13 countries with currency boards. Of these, six correspond to countries in the Eastern Caribbean Currency Union (Antigua and Barbuda, Dominica, Grenada, St Kitts and Nevis, St Lucia and St Vincent and the Grenadines), plus seven others: Bosnia and Herzegovina, Brunei Darussalam, Bulgaria, China-Hong Kong SAR, Djibouti, Estonia and Lithuania. Because all these countries are relatively small, currency boards are placed in a relatively unpopular category amongst potential exchange rate regimes.

There are two main reasons why countries have typically used currency boards. In some cases the currency board is more attractive than a common currency. For example, for the Eastern Caribbean countries mentioned above it seems relatively obvious they should use the US dollar as currency to maximize the benefits from a stable exchange rate arrangement with their almost sole trading partner. However, the currency board allows them to keep the exchange rate credibly fixed without giving up the seigniorage revenue of domestic currency. In other cases countries have resorted to a currency board as a way out of monetary and inflation chaos. Argentina's currency board experience in the 1990s and Bulgaria's currency board are appropriate examples. Even though, as we will see below, the evidence points to large trade benefits of currency boards, it is typically assumed that the main benefit of currency boards is as a tool to fight inflation.

The interest in and excitement about currency boards reflects both the need that countries have faced to solve either of the two problems mentioned above – currency integration without seigniorage cost and exiting from a high inflation situation – and the assessment made at the time of whether a currency board is the most efficient way to reach those objectives. Recent years have been unkind to currency boards on both counts. While the use of a currency board as a replacement for a common currency remains a valid motive, its effect as an anti-inflation device has become less relevant as inflation rates fell throughout the 1990s. In 2007 most countries exhibit single-digit inflation rates, and only a handful of exotic cases appear to have a monetary policy that is out of control. The high-inflation history of yesteryear has been critical to this improvement by fostering much stronger fiscal policies and monetary policies that are much freer from political pressures (both when central banks are independent and when they are not) and increasingly within an inflation targeting framework. As inflation has decreased, so have the benefits of a currency board, thus making it a relatively less attractive proposition. Furthermore, while before the demise of Argentina's currency board in early 2002 no currency board had been forced to end, the fact that Argentina's currency board came to an end in the midst of a major crisis (after enduring a long period of high interest rates) raised some questions as to how much credibility the regime actually bought. As a result, many countries have opted to jump directly all the way to dollarization (for example, El Salvador and Ecuador) or to pursue integration into a currency union (Slovenia) thus making currency boards lose ground even to alternative 'harder' exchange-rate commitments.

In spite of the recent drop in interest in this specific regime, nothing precludes a rise in interest again in the future, so a discussion of the specifics of currency boards remains useful. The best way to organize the discussion is to present the advantages of a currency board, then move to the disadvantages, and then attempt a synthesis.

Advantages of a Currency Board

The main advantage that is ascribed to a currency board is the credibility gains that it allows, helping deliver lower inflation and better fiscal results. The argument is simple: a currency board represents a strong commitment that if broken can have a large and costly effect on expectations. Because politicians fear this loss of credibility, the currency board, while in place, lowers inflation

expectations and inflation itself and should provide the incentives for an improvement in fiscal behaviour.

These predictions have been broadly borne out. On the inflation front Ghosh et al. (1998), drawing on a data-set for all IMF countries between 1970 and 1996, found that countries with currency boards delivered an inflation rate that was about four per cent lower, a sizable effect. This result has held up in later work (see for example Levy-Yeyati and Sturzenegger 2001; Kuttner and Posen 2001).

The record on fiscal discipline is also relatively favourable. Ghosh et al. (1998) and Culp et al. (1999) find that countries on currency boards tend to run tighter fiscal policies. Fatas and Rose (2000) also find that currency boards are associated with fiscal restraint (though, somewhat surprisingly, this restraint does not carry on to dollarized economies or those operating within the context of a common currency). Anecdotal evidence also seems to point in the same direction. In 2001, as Argentina's currency board was under fire, fiscal authorities implemented large budget adjustments in an attempt to strengthen the system.

Currency boards may also have an effect on trade as a result of the stability it induces on the exchange rate, an effect similar to the one that has been identified for countries that adopt a common currency with other countries. This exercise is specifically undertaken in Frankel and Rose (2002), who find that the effect of a currency board is a more than tripling of trade (in fact they find that the trade effects for currency boards and common currencies are statistically indistinguishable). Thus the trade motive for a currency board seems to be important. Added to the benefits of saving on seigniorage, it explains why currency boards may remain an attractive option for some small countries.

Disadvantages of a Currency Board

Four main arguments have been advanced against currency boards. First, the fact that it precludes monetary authorities from running an independent monetary policy and that the exchange rate cannot adjust in response to real shocks; second, that it may 'hide' underlying problems, leading to larger crises down the road; third, that it stimulates large currency mismatches in the portfolio structures of government and the private sector; and fourth, that it limits the ability of the central bank to act as a lender of last resort, thus hindering the possibility of developing a locally based financial sector.

The debate has focused mostly on whether alternative mechanisms and policies within the context of the currency board are available to deal with these problems. Let us review each of them briefly.

On the loss of monetary/exchange rate policy, the question is how relevant a loss this it. It can be argued that the idea of a currency board is indeed to limit the scope for an independent monetary policy, which had otherwise proven unable to contain high inflation. To the extent that inflation and fiscal policy improve, not much may be lost relative to the situation in which monetary policy merely induced inflation without any particular benefit in terms of macroeconomic stabilization. Thus, assessing whether this is a cost requires us to evaluate what the counterfactual is. Proponents of currency boards could argue that only countries where monetary policy serves no purpose choose currency boards as a commitment device.

Of course, if monetary policy *were* possible, the costs of doing without it may turn out to be particularly costly for currency boards. The case of Argentina helps illustrate why this should be so. Argentina had established a currency board with the dollar to quell inflation expectations in the early 1990s. Like any other emerging country, it was hurt by the rush out of emerging markets following Russia's default in 1998. This rush strongly appreciated the dollar, making Argentina's currency stronger exactly when the country needed it to weaken. The fact that currency boards require a strong anchor currency and that capital flows may strengthen these currencies when there is turmoil in emerging countries – thus moving the exchange rate exactly in the opposite direction to the one the country would have otherwise chosen – poses a problem for currency

boards during periods of high turbulence in international financial markets. Of course, as much as in the optimal currency area debate, how costly the loss of the monetary instrument is depends on the availability of alternative adjustment mechanisms: fiscal transfers, remittances, labour market mobility, or internal price flexibility, which may all operate as substitutes for the loss of monetary policy (the effectiveness of these alternative mechanisms may explain the different fates of Hong Kong's and Argentina's currency boards). Fiscal policy can also be used as a stabilizer that may substitute for the lack of exchange or monetary policy, though the ability of countries to use it seems relatively limited, particularly for those countries that opted for a currency board as a result of their poor fiscal policies. Some evidence for the fact that the lack of monetary policy may hurt is provided by Levy-Yeyati and Sturzenegger (2001), who compare the growth performance of hard pegs generally (including currency boards) with other regimes. They find that hard pegs trail floating regimes in growth performance (though not by more than pegs or intermediate regimes). However, this allows us to conclude that, in the end, the lack of policy responses may have a detrimental effect on overall economic performance.

The fact that currency boards may delay an adjustment has also been a cause of concern. Aizenman and Glick (2005) and Kuttner and Posen (2001) have both found that the harder and longer the peg, the larger are the depreciations upon exiting. This is to be expected, because the stronger the commitment, the fixed exchange rate spell will be typically longer, and only under more unfavourable conditions will the peg be abandoned, suggesting that an earlier adjustment may have been beneficial. This conclusion, however, should be treated with care because it fails to take into account the fact that this stringency also helps avoid many exits that later on would have turned out to be unnecessary.

The same caution should be used when evaluating the tendency of currency boards to foster the evolution of mismatches in government and private sector debt structures. The basic idea is that as long as the currency board holds countries develop a tendency to 'dollarize' their financial sectors (see Catao and Terrones 2000), with banks piling foreign currency deposits on their liability side, firms borrowing in dollars abroad and governments issuing debt in dollars. This is a problem because the asset side of these borrowers is in most cases linked mostly to the local economy, and thus, whether denominated in foreign currency or not, subject to currency risk in the event of a devaluation. This mismatch, however, is a double-edged sword. On the one hand it increases the commitment of the authorities to the peg (and this is why sometimes it is encouraged by the authorities as an additional credibility booster), but on the other it may also trigger large capital outflows in anticipation of a crisis. In the presence of large mismatches, agents would correctly anticipate a devaluation to produce a costly crisis, thus accelerating the run and the likelihood that the currency will sink. How these two factors play out during a crisis depends on the specifics of each individual country.

Finally, a currency board limits the ability of the central bank to operate as lender of last resort, particularly in the event of a bank run. This has been suggested as an explanation of why countries with currency boards quickly develop an international based banking system (typically with local institutions bought by foreign banks) which is better insured against runs at any specific location. Proponents of currency boards have suggested several alternatives to replace the central bank's function as lender of last resort with other mechanisms. Among these are the possibility of the government operating as lender of last resort, potentially by borrowing in dollars in times of need; the setting up of insurance schemes by which financial institutions buy in advance the access to funds in the context of a systemic liquidity run (these schemes were implemented by Mexico and Argentina); tighter capital and liquidity requirements on the banking sector; and the piling up of 'extra reserves' as far as possible. The first of these mechanisms is doubtful, as the government may have limited access to financing when it faces a crisis, and the others entail a cost. However, it may be said that some of these schemes have been implemented and used successfully. Specifically,

Argentina used its contingent credit line with private banks during its 2001 crisis and banks honoured their pledge at the time.

Where Does This Leave Us?

The conclusion is then that, as much as with currency unions, there seems to be a strong trade motive to set up a currency board. In fact, for a fiscally sound small country with the ability to conduct fiscal policy with some flexibility a currency board may be superior to a common currency as it allows the country to retain the seigniorage on its money stock. For larger middle-income countries a currency board has been pursued more as a way of improving credibility than anything else. While currency boards seem to have delivered, the Argentina case also suggests that their role in improving credibility cannot be taken fully for granted. If a currency board is implemented in times of easy access to international financial markets, fiscal discipline may be sidestepped and a fiscal and currency crisis may still occur at the end of the day. Additionally, policymakers should ask themselves if it makes sense to buy the credibility through a peg, or to buy it the hard way, day by day, implementing reasonable fiscal policies while maintaining some degree of flexibility in monetary policy. The successful experience since the mid-1990s of many countries with managed floating regimes and inflation targeting seems to point to this direction. If this trend continues, currency boards may become even rarer in the future.

See Also

- ► Currency Unions
- ► Dollarization

Bibliography

Aizenman, J., and R. Glick. 2005. *Pegged exchange rate regimes –a trap?* Working Paper No. 2006–07. Federal Reserve Bank of San Francisco.

Catao, L., and M. Terrones. 2000. *Determinants of dollarization: The banking side*, Working Paper No. 00/146. International Monetary Fund.

Culp, C., S. Hanke, and M. Miller. 1999. The case for an Indonesian currency board. *Journal of Applied Corporate Finance* 11: 57–65.

Fatas, A., and A.K. Rose . 2000. *Do monetary handcuffs restrain Leviathan? Fiscal policy in extreme exchange rate regimes*. Discussion Paper No. 2692. CEPR.

Frankel, J., and A. Rose. 2002. An estimate of the effect of common currencies on trade and income. *Quarterly Journal of Economics* 117: 437–466.

Ghosh A., A.-M. Gulde, and H.C. Wolf. 1998. *Currency boards: The ultimate fix?* Working Paper No. 98/8. International Monetary Fund.

Kuttner, K., and A. Posen. 2001. *Beyond bipolar: A three-dimensional assessment of monetary frameworks*, Working Paper No. 52. Oesterreichische Nationalbank.

Levy-Yeyati, E., and F. Sturzenegger. 2001. Exchange rate regimes and economic performance. *IMF Staff Papers* 47: 62–98.

Currency Competition

Stacey L. Schreft

Abstract

'Currency competition' means the virtually free entry of private-sector firms into the issuance of a currency. Such competition no longer exists, but interest in it revived in the 1970s as high inflation was attributed by some to governments' incentives to overissue their currencies to generate additional seigniorage. Competition was advocated as a potential remedy because it was thought to give issuers an incentive to protect the value of their currencies by limiting issuance.

Keywords

Bank Act 1844 (UK); Central banks; Currency competition; Currency School; Fiat money; Free banking; Hayek, F.; Inflation; Inside and outside money; Seigniorage; Suffolk Banking System

JEL Classifications

E42; E44

'Currency competition' refers to the free, or virtually free, entry of private-sector firms into the issuance of a circulating medium of exchange in lieu of a government monopoly on currency issue. Although there is little analytical basis for focusing on the private issuance of securities that circulate at the expense of those that do not, that is exactly the approach of the literature on currency competition and thus of this article.

The best real-world examples of currency competition come from periods, some lasting more than a century, in which countries allowed banks to operate relatively free from regulation. This freedom allowed, among other things, banks to issue paper notes. Shuler (1992) identified 66 countries as having free banking for some period in the 19th and 20th centuries, and all of them reportedly had multiple private-sector note issuers.

Today, there is no true private note issuance. Any privately issued notes are issued by banks that operate as agents of their respective central banks. Shuler (1992) attributed the demise of privately issued notes to several factors. One factor was a shift in attitudes about the need for and proper role of central banks. The view took hold that currency issuance could be destabilizing if left to the private sector, and governments nationalized currency issuance in their central banks. This was the case in England, for example, where the Currency School came to dominate and the Bank Act of 1844 eliminated private note issuance. Another major factor leading to government monopolies over currency issuance was the First World War and governments' need for additional sources of revenue. The ability to issue currency directly became very appealing.

By the 1970s, governments' monopoly on currency creation was raising its own concerns. These government issuers had an incentive to overissue to generate additional seigniorage revenue. When inflation began rising in the 1970s, some blamed this incentive to overproduce and called for denationalization of currency issuance. Friedrich Hayek (1990) was perhaps the most prominent proponent of a return to currency competition. Hayek argued, in the terms of today, that an equilibrium could exist with competitive issuance and that it would likely dominate the equilibrium arising when the government monopolizes currency issuance. The logic was that the demand for a privately issued currency depends in part on the currency's quality because such currencies are distinguishable. The more units of a currency supplied, the lower is the currency's value in exchange and thus its perceived quality and the public's demand for it. Competition would thus give issuers an incentive to protect the value of their currencies by not overissuing.

In considering what currency competition might look like, economists rediscovered the free banking periods, and a literature arose studying them. The first wave of that literature consisted of historical studies of free banking and private note issuance, although there were also a few theoretical models. Later, in the 1990s, the potential for new electronic means of payment, such as stored value and digital currencies for the Internet, led to another generation of research on currency competition, this time primarily theoretical.

Most discussions of currency competition, whether from a theoretical or an historical perspective, failed to distinguish inside money from outside money. Hellwig's work (1985) was an exception. Inside money is a claim that obligates its issuer to redeem or exchange the money for some specified monetary or nonmonetary object. Failure to do so, perhaps because of insufficient reserves held against the money, can result in a failure to fulfill that obligation and ultimately bankruptcy. The value of a privately issued inside money depends in part, then, on the likelihood of the issuer fulfilling its claim, and only in part on the value of using the money in exchange. Outside money is not a claim against the issuer or anyone else. The issuer makes no promise to redeem its currency at any time for anything of value. The value of a privately issued outside money derives solely from its value in exchange.

The experience in the US free banking era (1837–63) is an example of the importance of the claim that backs an inside money. Bank notes issued in the free banking era were supposed to be fully backed to guarantee the issuer's ability to redeem them, but often they were not. In some cases, no backing was held. Bank note reporters

kept track of the financial condition of issuers and of the prices at which notes were trading. Weber (2002) found that notes traded for one another at flexible exchange rates that often depended in part on the extent to which the notes were backed. When the public became aware that an issuer's notes lacked backing, the notes stopped circulating.

The distinction between inside and outside money is important for studying currency competition. Competition in outside note issuance is likely to divert fewer resources from consumption and production than competition in inside note issuance because there is no need to hold reserves against the outside notes. However, without reserves to back outside money, the money is likely to be overissued because of its near-zero marginal cost of production. Thus, the welfare gain from avoiding overissuing with an inside money must be balanced against the welfare loss from holding full or fractional reserves against such money.

Historical experience with outside money has almost always involved a single, government-issued fiat currency. The existing theoretical literature suggests why privately issued outside money is virtually never observed. In many different economic environments, economists have shown that there can be no equilibrium with competitive issuance of outside money if issuers cannot make binding commitments about the volume of notes they will issue. Taub (1985) and Bryant (1981) showed this in an overlapping-generations model. Ritter (1995) did so in a search model of money. In all cases, the argument is as follows, and similar to Hayek's. If issuing new money is costless, issuers cannot make binding commitments, and money has some positive value, then any private agent that issues notes will issue an unlimited quantity, driving the inflation rate to infinity and the real value of the money to zero. Rational agents would anticipate this ultimate outcome and be unwilling to hold the money at any earlier date. The inability to make binding commitments, coupled with a time inconsistency problem, is a key feature of this argument because issuers always want to believe they will constrain their note issuance, but when they need to they never have the incentive to do so.

A few models have gotten around this result. Klein (1974), for example, provided an early argument based on reputation formation for the existence of equilibria with free entry into private issuance. He argues that the monies of different issuers can be distinguishable by quality, so they can circulate at flexible exchange rates with one another. His discussion, however, blurs the distinction between inside and outside money.

In another example, Martin and Schreft (2006) showed that privately issued outside money can be valued if agents believe that all notes issued up to some threshold will be valued, but additional notes will be worthless. These beliefs create a discontinuity in the value of the marginal unit of currency. Because the value of a marginal unit of currency reaches zero for some finite supply, the limit argument no longer applies. Martin and Schreft derived their existence result in both an overlapping generations and a search-theoretic environment, though it should hold in any environment in which fiat currency could be valued. Interestingly, welfare is not necessarily greater with competitive issuance than with monopoly issuance and depends on the environment considered. In the search environment, neither competitive issuers nor a monopolist achieve the efficient quantity of money in the long run. In the overlapping-generations environment, the efficient allocation is achieved in finitely many periods if agents incur a cost of becoming money issuers. A monopoly issuer might achieve as desirable an allocation, but only if its actions are sufficiently constrained by agents' beliefs.

In contrast, the historical experience with inside money has involved multiple inside monies that are all convertible into some single dominant outside money. A modern literature on privately issued inside notes, largely attributable to Wallace and others, has considered this case. Cavalcanti and Wallace (1999a, b) studied a search-theoretic model with an exogenously given and indivisible outside money and inside money issued by private agents known as banks. To get the private money to be valued, they assumed that issuers who do not accept a note when presented with one face a stiff punishment: they lose the ability to issue notes and revert to autarky. This assumption is

reminiscent of the redemption requirements of successful systems for private inside currency issuance, like the Suffolk Banking System that operated in New England in the early 1800s. The authors found that, if the stock of outside money is sufficiently small, then the optimal mechanism has private notes issued and also redeemed on demand. Additionally, expected utility is greater in economies with inside money than only outside money because the set of implementable allocations is larger.

In the United States, at least, it is claimed that little currently prohibits private-sector issuance of outside currency in either paper or digital form. The laws prohibiting it have either expired or been repealed. It will be interesting to see if a resurgence of private issuance occurs.

See Also

- ▶ Fiat Money
- ▶ Free Banking Era
- ▶ Hayek, Friedrich August von (1899–1992)
- ▶ Inflation
- ▶ Inside and Outside Money

Bibliography

Bryant, J. 1981. The competitive provision of fiat money. *Journal of Banking & Finance* 5: 587–593.

de Cavalcanti, R.O., and N. Wallace. 1999a. A model of private bank-note issue. *Review of Economic Dynamics* 2: 104–136.

de Cavalcanti, R.O., and N. Wallace. 1999b. Inside and outside money as alternative media of exchange. *Journal of Money, Credit, and Banking* 31: 443–457.

Hayek, F. 1990. *Denationalisation of money: The argument refined*. 3 ed. London: Institute of Economic Affairs.

Hellwig, M. 1985. What do we know about currency competition? *Zeitschrift für Wirtschafts- und Sozialwissenschaften* 105: 565–588.

Klein, B. 1974. The competitive supply of money. *Journal of Money, Credit, and Banking* 6: 423–453.

Martin, A., and S. Schreft. 2006. Currency competition: A partial vindication of Hayek. *Journal of Monetary Economics* 53: 2085–2111.

Ritter, J. 1995. The transition from barter to fiat money. *American Economic Review* 85: 134–149.

Schreft, S. 1997. Looking forward: The role for government in regulating electronic cash. *Federal Reserve Bank of Kansas City Economic Review* (Fourth Quarter) 59–84.

Shuler, K. 1992. The world history of free banking: an overview. In *The experience of free banking*, ed. K. Dowd. New York: Routledge.

Taub, B. 1985. Private money with many suppliers. *Journal of Monetary Economics* 16: 195–208.

Weber, W. 2002. Banknote exchange rates in the antebellum United States. Working paper no. 623. Research Department, Federal Reserve Bank of Minneapolis.

C

Currency Crises

Graciela Laura Kaminsky

Abstract

This article describes models and empirical evidence on currency crises. The evidence from developed and developing countries indicates that crises are of different varieties. It also shows that crises do not occur in economies with sound fundamentals, with vulnerabilities far more widespread and profound in emerging economies. Vulnerabilities are associated with fiscal problems, loss of competitiveness and a deteriorating current account, external debt unsustainability, or problems in the financial sector – especially banks. Interestingly, those crises associated with bank fragility are the costliest in terms of output losses and loss of access to international capital markets.

Keywords

Budget deficits; Currency crises; Currency crisis models; Deposit insurance; European Monetary System; Exchange Rate Mechanism; Fixed exchange rates; Foreign-debt defaults; Imperfect information; International capital flows; International capital markets; Moral hazard; Regression tree analysis; Self-fulfilling currency crises; Sovereign defaults; Sticky prices; Sudden-stop currency crises

JEL Classifications

F3

A currency crisis occurs when investors flee from a currency en masse out of fear that it might be devalued. Currency crises are episodes characterized by sudden depreciations of the domestic currency, large losses of foreign exchange reserves of the central bank, and (or) sharp hikes in domestic interest rates.

There have been numerous currency crises since 1980. The so-called debt crisis erupted in 1982 following Mexico's default and devaluation in August. This crisis spread rapidly to all Latin American countries, and by the time it was over, most Latin American countries had devalued their currencies and defaulted on their foreign debts. The debt crisis was followed by a decade of negative growth and isolation from international capital markets. The output costs of this crisis were so large that the 1980s became known as the 'lost decade' for Latin America.

Crises are not just emerging-market phenomena. The 1990s opened with crises in industrial Europe – the European Monetary System (EMS) crises of 1992 and 1993. By the end of these crises, in the summer of 1993, the lira and the sterling had been driven from the Exchange Rate Mechanism (ERM); Finland, Norway, and Sweden had abandoned their unofficial peg to the European Currency Unit (ECU); the Spanish peseta, the Portuguese escudo and the Irish punt had devalued; and Europe's central bank governors and finance ministers had widened the ERM's intervention margins to ± 15 per cent from ± 2.25 per cent. Only then did the currency market stabilize.

Crises are hardy perennials. Within one year of the EMS crises, a currency crisis exploded in Mexico, with currency jitters spreading around the Latin American region. In 1997, it was Asia's turn. A new episode of currency turbulences started in July of that year with the depreciation of the Thai baht. Within a few days the crisis had spread to Indonesia, Korea, Malaysia and the Philippines. Turmoil in the foreign exchange market heightened in 1998 with the Russian default and devaluation in August. The Russian crisis spread around the world with speculative attacks in economies as far apart as South Africa, Brazil and Hong Kong. Currency crises have continued to erupt in the new millennium, with Argentina's crisis in December 2001 including the largest foreign-debt default in history.

The numerous financial crises that have ravaged emerging markets as well as mature economies have fuelled a continuous interest in developing models to explain why speculative attacks occur. Models are even catalogued into three generations. The first-generation models focus on the fiscal and monetary causes of crises. These models were mostly developed to explain the crises in Latin America in the 1960s and 1970s. In these models, unsustainable money-financed fiscal deficits lead to a persistent loss of international reserves and ultimately to a currency crash (see, for example, Krugman 1979).

The second-generation models aim at explaining the EMS crises of the early 1990s. These models focus on explaining why currency crises tend to happen in the midst of unemployment and loss of competitiveness. To explain these links, governments are modelled facing two targets: reducing inflation and keeping economic activity close to a given target. Fixed exchange rates may help in achieving the first goal but at the cost of a loss of competitiveness and a recession. With sticky prices, devaluations restore competitiveness and help in the elimination of unemployment, thus prompting the authorities to abandon the peg during recessions. Importantly, in this setting of counter-cyclical policies, the possibility of self-fulfilling crises becomes important, with even sustainable pegs being attacked and frequently broken (see, for example, Obstfeld 1994).

The next wave of currency crises, the Mexican crisis in 1994 and the Asian crisis in 1997, fuelled a new variety of models – also known as third-generation models – which focus on moral hazard and imperfect information. The emphasis here has been on 'excessive' booms and busts in international lending and asset price bubbles. These models also link currency and banking crises, sometimes known as the 'twin crises' (Kaminsky and Reinhart 1999). For example, Diaz-Alejandro (1985) and Velasco (1987) model difficulties in the banking sector as giving rise to a balance of payments crisis, arguing that, if central banks finance the bail-out of troubled financial

institutions by printing money, we have the classical story of a currency crash prompted by excessive money creation. Within the same theme, McKinnon and Pill (1995) examines the role of capital flows in an economy with an unregulated banking sector with deposit insurance and moral hazard problems of the banks. Capital inflows in such an environment can lead to over-lending cycles with consumption booms, real exchange rate appreciations, exaggerated current account deficits, and booms (and later busts) in stocks and property markets. Importantly, the excess lending during the boom makes banks more prone to a crisis when a recession unfolds. In turn, the fragile banking sector makes the task of defending the peg by hiking domestic interest rates more difficult and may lead to the eventual collapse of the domestic currency. Following the crisis in Argentina in 2001, the links between debt sustainability, sovereign defaults, and currency crises again attracted the attention of the economics profession. Finally, currency crises have also been linked to the erratic behaviour of international capital markets. For example, Calvo (1998) has brought to general attention the possibility of liquidity crises in emerging markets due to sudden reversals in capital flows, in large part triggered by developments in the world financial centres.

To summarize, all models suggest that currency crises erupt in fragile economies. Importantly, the three generations of models conclude that vulnerabilities come in different varieties. Still, the first attempts to study the vulnerabilities that precede crises have adopted 'the one size fits all' approach (see, for example, Frankel and Rose 1996; Kaminsky 1998). That is, the regressions estimated to predict crises include all possible indicators of vulnerability. These indicators include those related to sovereign defaults, such as high foreign debt levels, or indicators related to fiscal crises, such as government deficits, or even indicators related to crises of financial excesses, such as stock and real estate market booms and busts. In all cases, researchers impose the same functional form on all observations. When some indicators are not robustly linked to all crises, they tend to be discarded even when they may be of key importance for a subgroup of crises. Naturally, these methods leave many crises unpredicted and, furthermore, cannot capture the evolving nature of currency crises.

The next step in the empirical analysis of crises should be centred on whether crises are of different varieties. The first attempt in this direction is in Kaminsky (2006). In this article, a different methodology is used to allow for *ex ante* unknown varieties of currency crises. To identify the possible multiple varieties of crises, regression tree analysis is applied. This technique allows us to search for an unknown number of varieties of crises and of tranquil times using multiple indicators. This technique was also applied to growth by Durlauf and Johnson (1995).

Interestingly, this method catalogues crises into six classes:

1. *Crises with current account problems.* This variety is characterized by just one type of vulnerability, that of loss of competitiveness, that is, real exchange rate appreciations.
2. *Crises of financial excesses*. The fragilities are associated with booms in financial markets. In particular, they are identified as crises that are preceded by the acceleration in the growth rate of domestic credit and other monetary aggregates.
3. *Crises of sovereign debt problems.* These crises are characterized by fragilities associated with 'unsustainable' foreign debt.
4. *Crises with fiscal deficits.* This variety is just related to expansionary fiscal policy.
5. *Sudden-stop crises.* This type of crisis is only associated with reversals in capital flows triggered by sharp hikes in world interest rates, with no domestic vulnerabilities.
6. *Self-fulfilling crises*. This class of crises is not associated with any evident vulnerability, domestic or external.

These estimations allow us to answer four important questions about crises.

1. *Do crises occur in countries with sound fundamentals?* Even though this estimation allows for the identification of self-fulfilling crises (crises in economies with sound fundamentals), the results indicate that basically all

crises are preceded by domestic or external vulnerabilities. Only four per cent of the crises are unrelated to economic fragilities.

2. *How important are sudden reversals in capital flows in triggering crises?* While many have stressed that the erratic behaviour of international capital markets is the main culprit in emerging market currency crises, only two per cent of the crises in developing countries are just triggered by sudden-stop problems. While sudden-stop problems do occur, the reversals in capital flows mostly occur in the midst of multiple domestic vulnerabilities (see, Calvo et al. 2004).
3. *Are crises different in emerging economies?* Crises in emerging markets are preceded by far more domestic vulnerabilities than those in industrial countries. Overall, 86 per cent of the crises in emerging economies are crises with multiple domestic vulnerabilities, while economic fragility characterizes only 50 per cent of the crises in mature markets.
4. *Are some crises more costly than others?* It is a well-established fact that financial crises impose substantial costs on society. Many economists have emphasized the output losses associated with crises. But these are not the only costs of crises. In the aftermath of crises, most countries lose access to international capital markets, losing the ability to reduce the effect of adverse income shocks by borrowing in international capital markets. In most cases, countries have to run current account surpluses to pay back their debt. Finally, the magnitude of the speculative attack is itself important. For example, large depreciations may cause adverse balance sheet effects on firms and governments when their liabilities are denominated in foreign currencies. *Crises of financial excesses*, those also associated with banking crises – twin crisis episodes – are the costliest. Not only does the domestic currency depreciate the most, but also output losses are higher and the reversal of the current account deficit is attained via a dramatic fall in imports. In the aftermath of these crises, exports fail to grow even though the depreciations in this type of crises are massive. This evidence suggests that countries are even unable to attract trade credits to finance exports when their economies are mired in financial problems. In contrast, *self-fulfilling crises* and *sudden-stop crises* (but with no domestic vulnerabilities) have no adverse effects on the economies. Output (relative to trend) is unchanged or continues to grow in the aftermath of crises with no observed domestic fragility. In these crises, booming exports are at the heart of the recovery of the current account.

See Also

► Currency Crises Models

Bibliography

Calvo, G. 1998. Capital flows and capital-market crises: The simple economics of sudden stops. *Journal of Applied Economics* 1: 35–54.

Calvo, G., A. Izquierdo, and L. Mejia 2004. *On the empirics of sudden stops: The relevance of balance-sheet effects.* Working paper No. 10520. Cambridge, MA: NBER.

Diaz-Alejandro, C. 1985. Good-bye financial repression, hello financial crash. *Journal of Development Economics* 19: 1–24.

Durlauf, S., and P. Johnson. 1995. Multiple regimes and cross-country growth behavior. *Journal of Applied Econometrics* 10: 365–384.

Frankel, J., and A. Rose. 1996. Currency crises in emerging markets: An empirical treatment. *Journal of International Economics* 41: 351–366.

Kaminsky, G.L. 1998. *Currency and banking crises: The early warnings of distress*. International finance discussion papers No. 629. Board of Governors of the Federal Reserve System.

Kaminsky, G.L. 2006. Currency crises: Are they all the same? *Journal of International Money and Finance* 25: 503–527.

Kaminsky, G.L., and C. Reinhart. 1999. The twin crises: The causes of banking and balance-of-payments problems. *American Economic Review* 89: 473–500.

Krugman, P. 1979. A model of balance-of-payments crises. *Journal of Money, Credit, and Banking* 11: 311–325.

McKinnon, R.I., and H. Pill. 1995. Credible liberalizations and international capital flows: The 'overborrowing syndrome'. In *Financial deregulation and integration in East Asia*, ed. T. Ito and A.O. Krueger. Chicago: University of Chicago Press.

Obstfeld, M. 1994. The logic of currency crises. *Cahiers Economiques et Monétaires* 43: 189–213.

Obstfeld, M. 1996. Models of currency crises with self-fulfilling features. *European Economic Review* 40: 1037–1047.
Velasco, A. 1987. Financial and balance-of-payments crises. *Journal of Development Economics* 27: 263–283.

Currency Crises Models

Craig Burnside, Martin Eichenbaum and Sergio Rebelo

Abstract

Currency crises have occurred frequently in the post-war era. In this article we review the literature on the causes and consequences of currency crises. First- generation models attribute a central role to fiscal policy as a fundamental determinant of crises. Second-generation models emphasize the possibility of self-fulfilling speculative attacks and multiple equilibria. Third-generation models stress how financial fragility can lead to currency crises.

Keywords

Asian currency crisis (1997); Bank runs; Budget deficits; Consumer optimization; Credit risk; Currency crises; Currency crisis models; Exchange rate regimes; Financial crises; Fiscal theory of the price level; Fixed exchange rates; Government budget constraint; Government guarantees; Intertemporal budget constraint; Liquidity crises; Money demand functions; Money supply; Purchasing power parity; Seigniorage; Speculative attacks

JEL Classifications

D4; D10

There have been many currency crises during the post-war era (see Kaminsky and Reinhart 1999). A currency crisis is an episode in which the exchange rate depreciates substantially during a short period of time. There is an extensive literature on the causes and consequences of a currency crisis in a country with a fixed or heavily managed exchange rate. The models in this literature are often categorized as first-, second- or third-generation.

In first-generation models the collapse of a fixed exchange rate regime is caused by unsustainable fiscal policy. The classic first-generation models are those of Krugman (1979) and Flood and Garber (1984). These models are related to earlier work by Henderson and Salant (1978) on speculative attacks in the gold market. Important extensions of these early models incorporate consumer optimization and the government's intertemporal budget constraint into the analysis (see Obstfeld 1986; Calvo 1987; Drazen and Helpman 1987; Wijnbergen 1991). Flood and Marion (1999) provide a detailed review of first-generation models.

In a fixed exchange rate regime a government must fix the money supply in accordance with the fixed exchange rate. This requirement severely limits the government's ability to raise seigniorage revenue. A hallmark of first-generation models is that the government runs a persistent primary deficit. This deficit implies that the government must either deplete assets, such as foreign reserves, or borrow to finance the deficit. It is infeasible for the government to borrow or deplete reserves indefinitely. Therefore, in the absence of fiscal reforms, the government must eventually finance the deficit by printing money to raise seigniorage revenue. Since printing money is inconsistent with keeping the exchange rate fixed, first-generation models predict that the regime must collapse. The precise timing of its collapse depends on the details of the model.

The key ingredients of a first-generation model are its assumptions regarding purchasing power parity (PPP), the government budget constraint, the timing of deficits, the money demand function, the government's rule for abandoning the fixed exchange rate, and the post-crisis monetary policy. In the simplest first-generation models there is a single good whose domestic currency price is P_t and whose foreign currency price is 1. Let S_t denote the nominal exchange rate. PPP implies $P_t = S_t$. Suppose for simplicity that the

government has a constant ongoing primary deficit, δ. It finances this deficit by reducing its stock of foreign reserves, f_t, which can either evolve as a smooth function of time or jump discontinuously. In the former case, f_t evolves according to $\dot{f}_t = rf_t - \delta + \dot{M}_t/S_t$, where r is the real interest rate, M_t is the monetary base, and a dot over a variable denotes its derivative with respect to time. When foreign reserves change discontinuously, $\Delta f_i = \Delta(M_t/S_t)$. When $\delta > rf_0$ interest income from foreign assets will not be sufficient to finance the deficit.

To illustrate the key properties of first-generation models, we make three simplifying assumptions. First, money demand takes the Cagan (1956) form, $M_t = \theta P_t \exp[-\eta(r + \pi_t)]$, where $\theta > 0$ and $\pi_t = \dot{P}_t/P_t$ is the inflation rate. Second, the government abandons the fixed exchange rate regime when its foreign reserves are exhausted. Third, as soon as foreign reserves are exhausted, the government prints money at a constant rate μ to fully finance its deficit.

These assumptions imply that after the crisis the level of real balances, $m_t = M_t / P_t$, is constant and equal to $\overline{m} = \theta \exp[-\eta(r + \mu)]$. The post-crisis government budget constraint reduces to δ $\delta = \mu \overline{m}$. This equation determines μ. Let t^* denote the date at which foreign reserves are exhausted and the government abandons the fixed exchange rate regime. PPP implies $S_{t^*} = P_{t^*} = \overline{M}/\overline{m}$, where $\overline{M}$ is the monetary base the instant after date t^*. Under perfect foresight the exchange rate cannot jump discontinuously at t^* since such a jump would imply the presence of arbitrage opportunities. Given that the exchange rate must be a continuous function of time at t^*, $S_{t^*} = S$ and $\overline{M} = \overline{m}S$.

Prior to the crisis real balances are given by $m = \theta \exp(-\eta r)$. Therefore, at date t^* there is a sudden drop in real money demand from m to $\overline{m}$ implying that reserves drop discontinuously to zero at time $t^* : \Delta f_{t^*} = \overline{m} - m$. This is why the literature refers to t^* as the date of the speculative attack. Prior to the crisis the government's reserves fall at the rate $\dot{f}_t = rf_t - \delta$. The budget constraint implies that $t^* = \ln\{[\delta - r(m - \overline{m})]/(\delta - rf_0)\}/r$. While the collapse of the fixed exchange rate regime is inevitable, it does not generally occur at time zero unless $m - \overline{m} > f_0$.

A shortcoming of this type of first-generation model is that the timing of the speculative attack is deterministic and the exchange rate does not depreciate at the time of the attack. These shortcomings can be remedied by introducing shocks into the model, as in Flood and Garber (1984).

Early first-generation models predict that ongoing fiscal deficits, rising debt levels, or falling reserves precede the collapse of a fixed exchange rate regime. This prediction is inconsistent with the 1997 Asian currency crisis. This inconsistency led many observers to dismiss fiscal explanations of this crisis. However, Corsetti et al. (1999), Burnside et al. (2001a), and Lahiri and Végh (2003) show that bad news about prospective deficits can trigger a currency crisis. Under these circumstances a currency crisis will not be preceded by persistent fiscal deficits, rising debt levels, or falling reserves. These models assume that agents receive news that the banking sector is failing and that banks will be bailed out by the government. The government plans to finance, at least in part, the bank bailout by printing money beginning at some time in future. Burnside et al. (2001a) show that a currency crisis will occur before the government actually starts to print money. Therefore, in their model, a currency crisis is not preceded by movements in standard macroeconomic fundamentals, such as fiscal deficits and money growth. Burnside, Eichenbaum and Rebelo argue that their model accounts for the main characteristics of the Asian currency crisis.

This explanation of the Asian currency crisis stresses the link between future deficits and current movements in the exchange rate. This link is also stressed by Corsetti and Mackowiak (2006), Daniel (2001), and Dupor (2000), who use the fiscal theory of the price level to argue that prices and exchange rates jump in response to news about future deficits.

In first-generation models the government follows an exogenous rule to decide when to abandon the fixed exchange rate regime. In second-generation models the government maximizes an explicit objective function (see, for example, Obstfeld 1994, 1996). This

maximization problem dictates if and when the government will abandon the fixed exchange rate regime. Second-generation models generally exhibit multiple equilibria so that speculative attacks can occur because of self-fulfilling expectations. In Obstfeld's models (1994; 1996) the central bank minimizes a quadratic loss function that depends on inflation and on the deviation of output from its natural rate (see Barro and Gordon 1983, for a discussion of this type of loss function). The level of output is determined by an expectations-augmented Phillips curve. The government decides whether to keep the exchange rate fixed or not. Suppose agents expect the currency to devalue and inflation to ensue. If the government does not devalue then inflation will be unexpectedly low. As a consequence output will be below its natural rate. Therefore the government pays a high price, in terms of lost output, in order to defend the currency. If the costs associated with devaluing (lost reputation or inflation volatility) are sufficiently low, the government will rationalize agents' expectations. In contrast, if agents expect the exchange rate to remain fixed, it can be optimal for the government to validate agents' expectations if the output gains from an unexpected devaluation are not too large. Depending on the costs and benefits of the government's actions, and on agents' expectations, there can be more than one equilibrium. See Jeanne (2000) for a detailed survey of second-generation models.

Morris and Shin (1998) provide an important critique of models with self-fulfilling speculative attacks. They emphasize that standard second-generation models assume that fundamentals are common knowledge. Morris and Shin demonstrate that introducing a small amount of noise into agents' signals about fundamentals will lead to a unique equilibrium.

Many currency crises coincide with crises in the financial sector (Diaz- Alejandro 1985; Kaminsky and Reinhart 1999). This observation has motivated a literature that emphasizes the role of the financial sector in causing currency crises and propagating their effects. These third-generation models emphasize the balance- sheet effects associated with devaluations. The basic idea is that banks and firms in emerging market countries have explicit currency mismatches on their balance sheets because they borrow in foreign currency and lend in local currency. Banks and firms face credit risk because their income is related to the production of non-traded goods whose price, evaluated in foreign currency, falls after devaluations. Banks and firms are also exposed to liquidity shocks because they finance long-term projects with short-term borrowing. Eichengreen and Hausmann (1999) argue that currency mismatches are an inherent feature of emerging markets. In contrast, authors such as McKinnon and Pill (1996) and Burnside et al. (2001b) argue that, in the presence of government guarantees, it is optimal for banks and firms to expose themselves to currency risk.

Different third-generation models explore various mechanisms through which balance-sheet exposures may lead to a currency and banking crisis. In Burnside et al. (2004) government guarantees lead to the possibility of self-fulfilling speculative attacks. In Chang and Velasco (2001) liquidity exposure leads to the possibility of a Diamond and Dybvig (1983) style bank run. In Caballero and Krishnamurthy (2001) firms face a liquidity problem because they finance risky long-term projects with foreign loans but have access to limited amounts of internationally accepted collateral.

An important policy question is: what is the optimal nature of interest rate policy during and after a currency crisis? There has been relatively little formal work on this topic. Christiano et al. (2006) take an important first step in this direction. They argue that it is optimal to raise interest rates during a currency crisis and to lower them immediately thereafter. Studying optimal monetary policy in different models of currency crises remains an important area for future research.

See Also

- ▶ Currency Crises
- ▶ Fiscal Theory of the Price Level

Bibliography

Barro, R., and D. Gordon. 1983. A positive theory of monetary policy in a natural rate model. *Journal of Political Economy* 91: 589–610.

Burnside, C., M. Eichenbaum, and S. Rebelo. 2001a. Prospective deficits and the Asian currency crisis. *Journal of Political Economy* 109: 1155–1198.

Burnside, C., M. Eichenbaum, and S. Rebelo. 2001b. Hedging and financial fragility in fixed exchange rate regimes. *European Economic Review* 45: 1151–1193.

Burnside, C., M. Eichenbaum, and S. Rebelo. 2004. Government guarantees and self-fulfilling speculative attacks. *Journal of Economic Theory* 119: 31–63.

Caballero, R., and A. Krishnamurthy. 2001. International and domestic collateral constraints in a model of emerging market crises. *Journal of Monetary Economics* 48: 513–548.

Cagan, P. 1956. Monetary dynamics of hyperinflation. In *Studies in the quantity theory of money*, ed. M. Friedman. Chicago: University of Chicago Press.

Calvo, G. 1987. Balance of payments crises in a cash-in-advance economy. *Journal of Money Credit and Banking* 19: 19–32.

Chang, R., and A. Velasco. 2001. A model of financial crises in emerging markets. *Quarterly Journal of Economics* 116: 489–517.

Christiano, L., F. Braggion, and J. Roldos. 2006. The optimal monetary response to a financial crisis. Mimeo, Northwestern University.

Corsetti, G., and B. Mackowiak. 2006. Fiscal imbalances and the dynamics of currency crises. *European Economic Review* 50: 1317–1338.

Corsetti, G., P. Pesenti, and N. Roubini. 1999. What caused the Asian currency and financial crisis? *Japan and the World Economy* 11: 305–373.

Daniel, B. 2001. The fiscal theory of the price level in an open economy. *Journal of Monetary Economics* 48: 293–308.

Diamond, D., and P. Dybvig. 1983. Bank runs, deposit insurance and liquidity. *Journal of Political Economy* 91: 401–419.

Diaz-Alejandro, C. 1985. Good-bye financial repression, hello financial crash. *Journal of Development Economics* 19: 1–24.

Drazen, A., and E. Helpman. 1987. Stabilization with exchange rate management. *Quarterly Journal of Economics* 102: 835–855.

Dupor, W. 2000. Exchange rates and the fiscal theory of the price level. *Journal of Monetary Economics* 45: 613–630.

Eichengreen, B., and R. Hausmann. 1999. Exchange rates and financial fragility. In *New challenges for monetary policy: A symposium sponsored by the Federal Reserve Bank of Kansas City*. Kansas City: Federal Reserve Bank of Kansas City.

Flood, R., and P. Garber. 1984. Collapsing exchange rate regimes: Some linear examples. *Journal of International Economics* 17: 1–13.

Flood, R., and N. Marion. 1999. Perspectives on the recent currency crisis literature. *International Journal of Finance and Economics* 4: 1–26.

Henderson, D., and S. Salant. 1978. Market anticipations of government policies and the price of gold. *Journal of Political Economy* 86: 627–648.

Jeanne, O. 2000. Currency crises: A perspective on recent theoretical developments. Special Papers in International Economics, No. 20, International Finance Section, Princeton University.

Kaminsky, G., and C. Reinhart. 1999. The twin crises: The causes of banking and balance-of-payments problems. *American Economic Review* 89: 473–500.

Krugman, P. 1979. A model of balance of payments crises. *Journal of Money, Credit and Banking* 11: 311–325.

Lahiri, A., and C. Végh. 2003. Delaying the inevitable: Interest rate defense and BOP crises. *Journal of Political Economy* 111: 404–424.

McKinnon, R., and H. Pill. 1996. Credible liberalizations and international capital flows: The overborrowing syndrome. In *Financial deregulation and integration in East Asia*, ed. T. Ito and A. Krueger. Chicago: University of Chicago Press.

Morris, S., and H. Shin. 1998. Unique equilibrium in a model of self-fulfilling currency attacks. *American Economic Review* 88: 587–597.

Obstfeld, M. 1986. Speculative attack and the external constraint in a maximizing model of the balance of payments. *Canadian Journal of Economics* 29: 1–20.

Obstfeld, M. 1994. The logic of currency crises. *Cahiers Economiques et Monétaires* 43: 189–213.

Obstfeld, M. 1996. Models of currency crises with self-fulfilling features. *European Economic Review* 40: 1037–1047.

van Wijnbergen, S. 1991. Fiscal deficits, exchange rate crises and inflation. *Review of Economic Studies* 58: 81–92.

Currency Unions

Andrew K. Rose

Abstract

This article reviews currency unions, that is, groups of countries that use a common money. There are a large number of such monetary unions in both the industrial and the developing worlds. I review both the theoretical reasons why countries choose to belong to currency unions and the empirical performance of these unions.

I thank Steven Durlauf for helpful comments.

Keywords

Asymmetric shocks; Automatic stabilizers; Business cycles; Capital mobility; Countercyclical fiscal policy; Currency unions; Diversification; Economic and Monetary Union (EMU); Exchange rate regimes; Inflation; Labour mobility; Latin Monetary Union; Monetary unions *see* currency unions; Mundell, R; Optimal currency area; Output volatility; Progressive taxation; Rigidities; Scandinavian Monetary Union; Size of nations

JEL Classifications

F3

Currency unions (also known as monetary unions) are groups of countries that share a single money. Currency unions are unusual, since most countries have their own currency. For instance, the United States, Japan and the United Kingdom all have their own monies. But a reasonable number of countries participate in currency unions, and their importance is growing. In May 2005, 52 of the 184 IMF members participated in currency unions.

Currency Unions Present and Past

Currency unions commonly come about when a small or poor country unilaterally adopts the money of a larger, richer 'anchor' country. For instance, a number of countries currently use the US dollar, including Panama, El Salvador, Ecuador, and a number of smaller countries and dependencies in the Caribbean and Pacific. Swaziland, Lesotho and Namibia all use the South African rand. Both the Australian and New Zealand dollars are used by a number of countries in the Pacific; Liechtenstein uses the Swiss franc; and so forth. In the past, a number of countries have used the currency of their colonizer; over 50 countries and dependencies have used the British pound sterling at one time or another. Cases like this are known as official dollarization (unofficial dollarization occurs when the currency of a foreign country circulates widely but is not formally the national currency). In such cases, the small country essentially relinquishes its right to sovereign monetary policy. It loses its ability to independently influence its exchange and interest rates; these are determined by the anchor country, typically on the basis of the interests of the anchor.

There are also a number of multilateral currency unions between countries of more or less equal size and wealth. For instance, the East Caribbean dollar circulates in Anguilla, Antigua and Barbuda, Dominica, Grenada, Montserrat, Saint Kitts and Nevis, Saint Lucia, and Saint Vincent and the Grenadines. The Central Bank of the West African States circulates the Communautéfrançaise d'Afrique (CFA) franc in Benin, Burkina Faso, Côte d'Ivoire, Guinea-Bissau, Mali, Niger, Senegal, and Togo. The Bank of the Central African States circulates a slightly different CFA franc in Cameroon, the Central African Republic, Chad, Republic of Congo, Equatorial Guinea, and Gabon.

The largest and most important currency union is the Economic and Monetary Union of the European Union (EMU). EMU technically began on 1 January 1999, although the euro was physically introduced only three years later. Twelve countries are formally members of EMU: Austria, Belgium, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, the Netherlands, Portugal, and Spain. (A number of smaller European territories and French dependencies also use the euro.) These countries jointly determine monetary policy for EMU through the international European Central Bank. The number of members in EMU is expected to grow with time, especially as countries that acceded to the European Union in 2004 become eligible for EMU entry. However, both Sweden and Denmark have rejected membership in referenda, and the euro remains unpopular in the UK.

While a number of currency unions currently exist, many have not survived. The Latin Monetary Union began in 1865 when France, Belgium, Italy and Switzerland (later joined by Greece, Romania, and others) adopted common regulations for their individual currencies to encourage the free international flow of money. This essentially amounted to a commitment to mint silver

and gold coins to uniform specifications, but without other restrictions on monetary policy. The union effectively ended with the onset of the First World War. The war also ended the Scandinavian Monetary Union which Denmark, Norway, and Sweden began in 1873. The economic union between Belgium and Luxembourg that began in 1921 has been absorbed into EMU. Multilateral currency unions in East Africa, Central Africa, West Africa, South Asia, South-East Asia, and the Caribbean have also disappeared.

Theory: Why Should Countries Enter Currency Union?

Historically, most countries have had their own moneys. There seems to be a tight connection between national identity and national money; a country's money is a potent symbol of sovereignty. Still, some countries have entered into currency union. Why? Economists have theorized about the potential economic benefits of currency union which can, in certain circumstances, overwhelm the perceived political costs.

Like all other monetary regimes, currency unions are fully compatible with Robert Mundell's (1968) celebrated 'Trilemma' or 'Incompatible Trinity'. A country would like its monetary regime to deliver three desirable goals that turn out to be mutually exclusive: domestic monetary sovereignty, capital mobility, and exchange rate stability. Currently, large rich countries like the United States, Japan and the UK have domestic monetary sovereignty and open capital markets but have floating exchange rates. By way of contrast, members of a currency union essentially relinquish the first objective (monetary independence) in exchange for the latter benefits (capital mobility and stable exchange rates). Indeed, some economists think of currency unions as simply extreme forms of fixed exchange rates, with all the associated pros and cons. Countries inside currency union receive more microeconomic benefits than they would from a fixed exchange rate, since sharing a single money leads to deeper integration of real and financial markets. On the other hand, a country can devalue or float the exchange rate more easily than it can leave a currency union. Still, this is an unsatisfying theoretical approach the issue of currency unions. It does not address to the vital question: what is the optimal size of a currency union? If the right size for a currency union is not necessarily the country, how should we tackle the problem?

The theoretical analysis of currency unions began with a seminal paper by Mundell (1961). Mundell's analysis answered the question: what is the appropriate domain for a currency? Mundell briefly argued there are advantages to regions that use a common money. In particular, currency union facilitates international trade; a single medium of exchange reduces transactions costs, as does a common unit of account. However, a common currency can also cause problems in the dual presence of asymmetric shocks and nominal rigidities (in prices and wages). Suppose demand shifts from Western to Eastern goods. The increase in demand for Western output results in inflationary pressures there, while East goes into recession. Mundell argued that, if unemployed labour could move freely from East to relieve inflationary pressures in West, the two problems could be resolved simultaneously. However, in the absence of labour mobility, the asymmetric shock could be better handled by allowing the Western currency to appreciate. But in order for this to happen, both East and West must have their own monies! Mundell concluded that the optimal currency area was the area within which labour is mobile; regions of labour mobility should have their own currencies.

Two other classic contributions to the theory of optimal currency areas are worthy of note. McKinnon (1963) examined the effects of country size on currency unions; he concluded that smaller countries tend to be more open and have fewer nominal rigidities, making them better candidates for currency union. Kenen (1969) considered the effects of the economy's degree of diversification, and argued that more diversification resulted in fewer asymmetric shocks, and accordingly fewer benefits from national monetary policy.

The key focus of Mundell's theoretical optimum currency area framework – the adjustment to

asymmetric shocks – has stood the test of time well. The ability of a region to respond to such shocks is viewed as a critical part of a sustainable and desirable currency union. Still, hardly anyone now takes the narrow specifics of Mundell's original article seriously. In particular, Mundell's conclusion that the optimum currency area is a region of labour mobility is no longer widely believed. The problem of asymmetric business cycles that Mundell described is intrinsically a problem of. . . business cycles. The costs of shifting labour are high almost everywhere in the world, which is why labour moves only slowly, even within countries with relatively flexible labour markets like the United States. Accordingly, most economists are uncomfortable thinking that labour could or should shift in response to the shocks and propagation mechanisms that cause business cycles. After all, the nominal rigidities that are responsible for business cycles do not last for ever. Thus, Mundell's idea of labour mobility is no longer viewed as a viable adjustment mechanism. (This conclusion is tempered if one believes that real shocks cause business cycles without nominal rigidities.)

Still, there are other ways to share the risks of, or adjust to, asymmetric shocks, and much of the relevant work has incorporated these other mechanisms. Mundell originally ignored capital mobility. But private capital markets can, in principle, spread shocks internationally if investors diversify across regions or sectors. However, more attention has been paid to the public sector, since a federal system of taxes and transfers may be an efficient way to spread risks across regions. To continue with the East and West example, a progressive federal tax structure reduces inflationary Western pressures, and allows benefits to be paid to the unemployed in the East. Both regions suffer less macroeconomic volatility with such automatic stabilizers in place. The most controversial adjustment mechanism is counter-cyclical fiscal policy. In response to an asymmetric shock, regions that are free and capable of deploying discretionary fiscal policy can uses changes in taxes and government spending to respond to asymmetric shocks, even within the monetary confines of a currency union. More generally, mechanisms to handle asymmetric shocks are still an integral part of the theory of currency unions.

Mundell originally thought the great benefit of currency union was the facilitation of trade since money is a convenience that lowers transactions costs. But suppose that countries produce moneys of different qualities. Argentina has gone through five currencies since 1970; high Argentine inflation results in a low convenience value for Argentine money. Suppose Argentina decides to give up on a national money altogether and enter into a currency union with a foreign producer of higher-quality money: the United States, say. Argentina will surely experience different shocks from the United States, and these shocks have to be handled. Perhaps then Argentina should enter a currency union with a country with more similar shocks? The problem is that the most obvious contender, Brazil, also has a history of monetary incompetence. The larger point is that a low-quality domestic monetary authority increases a country's willingness to enter currency union, as does the availability of high-quality foreign money. Alesina and Barro (2002) provide an elegant model that incorporates such features. In their model, countries enter currency unions with neighbours in order to facilitate trade, so long as the neighbours possess monetary institutions of quality. Lower inflation and reduced transactions costs of trade provide gains, while the inability to respond to idiosyncratic asymmetric shocks generates losses.

Empirics: What Do We Know in Practice About Currency Unions?

During the run-up to EMU, a considerable empirical literature developed that quantified different aspects of optimal currency areas. Much attention was paid to estimating the synchronization of business cycles for potential EMU candidates; Bayoumi and Eichengreen (1992) was the first important paper. The tradition has since been generalized to more countries by Alesina et al. (2002), who characterized co-movements in prices as well as output. Frankel and Rose (1998) showed that the intensity of trade had a strong positive effect

on business cycle synchronization; that is, the optimum currency area criteria are jointly endogenous. If currency union lowers the transactions costs of trade and thus leads to an increase in trade, it may also thereby reduce the asymmetries in business cycles; areas that do not look like currency unions ex *ante* may do so *ex post.* Bayoumi and Eichengreen (1998) successfully link optimum currency area criteria (principally the asymmetry of business cycle shocks) to exchange rate volatility and intervention, and show that a number of features of the optimum currency area theory appear in practice, even for countries not in currency unions.

Somewhat curiously, little work was done to analyse actual currency unions until around 2000. This is probably because the currency unions that preceded EMU consisted mostly of small or poor countries, which were viewed as irrelevant for EMU. But this gap in the literature implicitly allowed economists to focus their attention on the costs of currency union, which tend to be macroeconomic in nature (resulting from the absence of national monetary policy as a tool to stabilize business cycles). As Mundell clearly pointed out, there are also benefits from a currency union, mostly microeconomic in nature. Fewer monies mean lower transactions costs for trade, and thus higher welfare. An unresolved issue of importance is the size of the benefits that stem from currency union. There is evidence that currency unions have been associated with increased trade in goods, though its size is much disputed. Using data on pre-EMU currency unions (such as the CFA franc zone), Rose (2000) first estimated the effect of currency union on trade, and found it to result in an implausibly high tripling of trade. This finding and the intrinsic interest of EMU have resulted in a literature that has almost universally found smaller estimates, which are yet of considerable economic size. Rose and Stanley (2005) provide a quantitative survey that concludes that currency union increases trade by between 30 and 90 per cent. Engel and Rose (2002) examine other macroeconomic aspects of pre-EMU currency unions, and find that currency union members are more integrated than countries with their own monies, but less integrated than regions within a single country. Edwards and Magendzo (2003) compare inflation, output growth and output volatility in countries inside currency unions and those outside them, and find that currency unions have lower inflation and higher output volatility than countries with their own currencies.

Areas of Ignorance

The impact of currency union on financial markets is not something that is currently well understood. Yet this is an area of great interest, since currency union might result in deeper financial integration – or it might not. It is clearly of concern to the British government, which has made the financial effects one of its five tests for EMU entry (see HM Treasury 2003).

More generally, Europe's experiment with currency union is still young. It is simply too early to know whether EMU has resulted in substantial changes in the real economy, financial or labour markets, or political economy. As the data trickles in, most expect a continuing reassessment of currency unions in theory and especially practice.

See Also

► Mundell, Robert (Born 1932)

Bibliography

Alesina, A., and R. Barro. 2002. Currency unions. *Quarterly Journal of Economics* 117: 409–436.

Alesina, A., R. Barro, and S. Tenreyro. 2002. Optimal currency areas. In *NBER macroeconomics annual 2002*, ed. M. Gertler and K. Rogoff. Cambridge, MA: MIT Press.

Bayoumi, T., and B. Eichengreen. 1992. Shocking aspects of European monetary unification. In *The transition to economic and monetary union in Europe*, ed. F. Torres and F. Giavazzi. New York: Cambridge University Press.

Bayoumi, T., and B. Eichengreen. 1998. Exchange rate volatility and intervention: Implications of the theory of optimum currency areas. *Journal of International Economics* 45: 191–209.

Edwards, S., and I. Magendzo. 2003. A currency of one's own? An empirical investigation on dollarization and

independent currency unions. Working Paper No. 9514. Cambridge, MA: NBER.
Engel, C., and A. Rose. 2002. Currency unions and international integration. *Journal of Money Credit and Banking* 34: 1067–1089.
Frankel, J., and A. Rose. 1998. The endogeneity of the optimum currency area criteria. *Economic Journal* 108: 1009–1025.
HM Treasury. 2003. *UK membership of the single currency: EMU studies.* Online. http://www.hm-treasury.gov.uk/documents/international_issues/the_euro/assessment/studies/euro_assess03_studindex.cfm. Accessed 25 Mar 2006.
Kenen, P. 1969. The theory of optimum currency areas: An eclectic view. In *Monetary problems of the international economy*, ed. R. Mundell and A. Swoboda. Chicago: University of Chicago Press.
McKinnon, R. 1963. Optimum currency areas. *American Economic Review* 53: 717–724.
Mundell, R. 1961. A theory of optimum currency areas. *American Economic Review* 51: 657–665.
Mundell, R. 1968. *International economics*. New York: Macmillan.
Rose, A. 2000. One money, one market: Estimating the effect of common currencies on trade. *Economic Policy* 30: 7–46.
Rose, A., and T. Stanley. 2005. A meta-analysis of the effect of common currencies on international trade. *Journal of Economic Surveys* 19: 347–365.

Currie, Lauchlin (1902–1993)

Roger Sandilands

Abstract

At Harvard in the early 1930s Currie pioneered a monetary diagnosis of the 1929–32 collapse and placed blame on the Federal Reserve Board. As a prominent New Dealer at the Fed during 1934–9 he urged contra-cyclical monetary and fiscal activism. During 1939–45 he worked in Washington as President Roosevelt's economic adviser. After heading a World Bank mission to Colombia in 1949 he spent 40 years advising on national development there. He emphasized urban housing as a leading sector, based on an innovative housing finance system, and extended Allyn Young's ideas on macroeconomic increasing returns and endogenous growth.

Keywords

Bretton Woods conference; Commercial loan theory of banking; Credit; Currie, L.; Endogenous growth; Federal Reserve System; Great Depression; Hansen, A.; Income velocity of money; Increasing returns; Land tax; Monetary and financial forces in the Great Depression; Plan of the Four Strategies (Colombia); Quantity of money; Reserve requirement; Rostow, W.; Urban planning; Viner, J.; White, H.; Young, A.

JEL Classifications

B31

Lauchlin Currie was born on 8 October 1902 in West Dublin, Nova Scotia, and died in Bogotá, Colombia, on 23 December 1993 after an unusually long and varied career as an academic economist and top-level policy adviser. After two years at St Francis Xavier University, Nova Scotia, 1920–2, he moved to the London School of Economics (LSE), where his teachers included Edwin Cannan, Hugh Dalton, A. L. Bowley, R. H. Tawney and Harold Laski. In 1925 he obtained his BsC and moved to Harvard, where the chief inspiration for his Ph.D. thesis, 'Bank Assets and Banking Theory' (January 1931), was Allyn Abbott Young. However, when Young moved to the LSE in 1927 his formal supervisor was John H. Williams.

He remained at Harvard until 1934 as teaching assistant to Williams, Ralph Hawtrey and Joseph Schumpeter. His Ph.D. thesis attacked the 'commercial loan' or 'needs of trade' theory of banking by showing that it was not only unsound in theory but had been more honoured in the breach than the observance – until its disastrous influence on monetary policy in the late 1920s and early 1930s.

In a January 1932 memorandum, Currie, Harry Dexter White and Paul Theodore Ellsworth presented a radical anti-depression programme (see Laidler and Sandilands 2002). In keeping with their explanation of the contraction as due to a collapsing money supply, they urged vigorous open-market operations and deficit spending

financed by money creation. This memorandum was part of an early Harvard influence (through Young, Hawtrey, Williams and Currie; see Laidler 1999) on what had been claimed as a unique Chicago monetary tradition.

In Currie (1933a) he showed the hopeless confusion that resulted from the ambiguity of the word 'credit'. He stressed control over the quantity of money (defined as cash plus demand deposits, for which there had been no estimates until Currie published a series in 1934) rather than the quantity or quality of credit or loans. He also computed the first estimate of the income velocity of money in the United States (Currie 1933b), with an explanation of its cyclical variations.

His 'The Failure of Monetary Policy to Prevent the Depression of 1929–32' (1934a) fully anticipated Milton Friedman and Anna Schwartz's (1963) diagnosis of this period. He argued that apart from the stock market there were none of the traditional signs of a boom in the 1920s. Tight monetary policies had been ineffectual in checking the rise in stock prices but only too effective in contributing to the decline in building activity and the pressure on foreign countries that preceded the Depression.

He also demonstrated the perverse elasticity of money in the business cycle due to differences in reserve requirements for different classes of bank and bank deposit (1934b). In the face of the banks' reserve losses in 1929–32 and their abhorrence of heavy indebtedness to the reserve banks, the administration's policy was 'one of almost complete passivity and quiescence', so the self-generating forces of the Depression continued unchecked.

In 1934 Jacob Viner recruited him to the 'freshman brain trust' at the US Treasury where he developed a blueprint for a system of 100 per cent reserves against demand deposits, to break the link between the lending and the creation of money and to strengthen central bank control (see Phillips 1995). Later that year Marriner Eccles, the new governor of the Federal Reserve Board, hired Currie as his top adviser, from 1934 to 1939. (Many of his memoranda to Eccles are published in Sandilands 2004.)

At the Fed Currie drafted what became the 1935 Banking Act that gave the Fed increased powers to raise reserve requirements. In 1936–7 these powers were used, 'as a precautionary measure', to reduce the huge build-up of banks' excess reserves. This has been widely blamed for the sharp recession of 1937–8, a view Currie consistently rejected (1938). Instead, he invoked his newly constructed 'net federal income-creating expenditure series' (1935; and see Sweezy 1972) to show the strategic role of fiscal policy in complementing monetary policy to revive an acutely depressed economy. In November 1937 he had a four-hour meeting with President Roosevelt to explain that the recession was due to sharp fiscal contraction and that balancing the budget was not the way to restore business confidence. He insisted on the need for better coordination of monetary and fiscal policy. In May 1939 the rationale for this was explained in theoretical and statistical detail by Currie and Alvin Hansen (respectively 'Mr Inside' and 'Mr Outside', according to Tobin 1976), in joint testimony before the Temporary National Economic Committee.

From 1939 to 1945, Currie was President Roosevelt's special adviser on economic affairs in the White House. He was also in charge of lend-lease to China, 1941–3, and ran the Foreign Economic Administration, 1943–4. In early 1945 he headed a tripartite (United States, British and French) mission to Bern to persuade the Swiss to freeze Nazi bank balances and stop shipments of German supplies through Switzerland to the Italian front. He was also closely involved in loan negotiations with British and Soviet allies and in preparations for the 1944 Bretton Woods conference (staged primarily by his friend Harry White).

After the war it was alleged by Elizabeth Bentley, an ex-Soviet agent, that Currie and White had participated in Soviet espionage. Though she had never met them herself, she claimed they had passed information to other Washington economists who were abetting her own espionage, and that they probably knew this. White and Currie were heavily involved in official wartime cooperation with the Soviets, but Bentley put a sinister interpretation on these activities. They appeared

together before the House Committee on Un-American Activities in August 1948 to rebut Bentley's charges. Their testimony satisfied the Committee at that time, though the strain contributed to the fatal heart attack that White suffered three days after the hearing.

No charges were laid against Currie, and in 1949 he headed a major World Bank survey of Colombia. In 1950 the Colombians invited him to return to Bogotá, where he remained for most of the next 40 years as a top presidential adviser. He has been falsely accused of fleeing the United States to avoid charges of disloyalty. In fact in December 1952 he was a witness before a grand jury in New York investigating Owen Lattimore's role in the famous *Amerasia* case that involved the publication of secret State Department documents by that magazine, though his next visit to the United States was not until 1961 when he had a meeting in the White House with Walt Rostow, then President Kennedy's National Security Adviser, to discuss a development plan for Colombia.

By that time Currie had assumed Colombian citizenship (personally conferred on him by President Alberto Lleras in 1958), partly because in 1954 the US government had refused to renew his passport, ostensibly because he was only a naturalized US citizen and was now residing abroad. However, the reality was probably connected with the then secret 'Venona' project that had deciphered wartime Soviet cables that mentioned Currie. The related cases of Currie and White are discussed in Sandilands (2000) and Boughton and Sandilands (2003), where it is shown that the evidence against them is far from conclusive. After reading the latter paper, Major-General Julius Kobyakov, deputy director of the KGB's American desk in the late 1980s, wrote to the present writer on 22 December 2003 to confirm our conclusions. After extensive archival research on Soviet intelligence in the 1930s and 1940s he found that

> there was nothing in [Currie's] file to suggest that he had ever wittingly collaborated with the Soviet intelligence. . . However, in the spirit of machismo, many people claimed that we had an 'agent' in the White House. Among the members of my profession there is a sacramental question: 'Does he know that he is our agent?' There is very strong indication that neither Currie nor White knew that.

There were two breaks to Currie's advisory and academic work in Colombia: during a military dictatorship, 1953–8, he retired to develop a prize-winning herd of Holstein cattle; and from 1966 to 1971 he was a professor at Michigan State (1966), Simon Fraser (1967–8 and 1969–71), Glasgow (1968–9), and Oxford (1969) universities. He returned permanently to Colombia in 1971 at the behest of President Misael Pastrana to prepare a national plan of development known as the Plan of the Four Strategies, with a focus on urban housing and export diversification. The plan was implemented and the institutions that were established in support of the plan played a major role in accelerating Colombia's urbanization.

He remained as chief economist at the National Planning Department for ten years, 1971–81, followed by 12 years at the Colombian Institute of Savings and Housing until his death in 1993. There he defended the unique index-linked housing finance system (based on 'units of constant purchasing power' for both savers and borrowers) that he had established in 1972. The system thus continued to boost Colombia's growth rate and urban employment opportunities year by year. Currie was also a top adviser on urban planning, and played a major part in the first United Nations Habitat conference in Vancouver in 1976. His 'cities-within-the-city' urban design and financing proposals (including the public recapture of land's socially created 'valorización', or 'unearned land value increments', as cities grow) were elaborated in *Taming the Megalopolis* (1976). To the time of his death he was a regular teacher at the National University of Colombia, Javeriana University, and the University of the Andes, and continued to publish widely (a comprehensive bibliography is in Sandilands 1990, reviewed by Charles Kindleberger 1991). His writings and policy advice were heavily influenced by his old Harvard mentor, Allyn Young. Notable is his posthumous (1997) paper that offers a unique macroeconomic interpretation of Youngian increasing returns and the endogenous nature of self-sustaining growth.

See Also

- ▶ Development Economics
- ▶ Endogenous Growth Theory
- ▶ Federal Reserve System
- ▶ Great Depression, Monetary and Financial Forces in
- ▶ Hansen, Alvin (1887–1975)
- ▶ Housing Supply
- ▶ Monetary Policy, History of
- ▶ Urbanization
- ▶ Viner, Jacob (1892–1970)
- ▶ Young, Allyn Abbott (1876–1929)

Selected Works

1931. Bank assets and banking theory. Ph.D. thesis, Harvard University.
1932. (With P. Ellsworth and H. White.) Memorandum on anti-depression policy. *History of Political Economy* 34(2002): 533–552.
1933a. The treatment of credit in contemporary monetary theory. *Journal of Political Economy* 41: 509–525.
1933b. Money, gold and incomes in the United States, 1921–32. *Quarterly Journal of Economics* 48: 77–95.
1934a. The failure of monetary policy to prevent the Depression of 1929–32. *Journal of Political Economy* 42: 145–77. Reprinted in *Landmarks in political economy,* ed. E. Hamilton, H. Johnson and A. Rees. Chicago: University of Chicago Press, 1962.
1934b. *The supply and control of money in the United States.* Cambridge, MA: Harvard University Press.
1935. Comments and observations on 'Federal Income-Increasing Expenditures, 1933–35'. *History of Political Economy* 10(1978): 507–548.
1938. Causes of the recession. *History of Political Economy* 12(1980): 303–335.
1976. *Taming the megalopolis: A design for urban growth.* Oxford: Pergamon Press.
1997. Implications of an endogenous theory of growth in Allyn Young's macroeconomic concept of increasing returns. *History of Political Economy* 29, 414–443.

Bibliography

Boughton, J., and R. Sandilands. 2003. Politics and the attack on FDR's economists: From Grand Alliance to Cold War. *Intelligence and National Security* 18 (3): 73–99.
Friedman, M., and A. Schwartz. 1963. *A monetary history of the United States, 1867–1960*. Princeton: Princeton University Press.
Kindleberger, C. 1991. Review of Roger J. Sandilands, *The life and political economy of Lauchlin Currie. Journal of Political Economy* 99: 1119–1122.
Laidler, D. 1999. *Fabricating the Keynesian revolution*. Cambridge: Cambridge University Press.
Laidler, D., and R. Sandilands. 2002. An early Harvard *Memorandum* on anti-depression policies. *History of Political Economy* 34: 515–532.
Phillips, J. 1995. *The Chicago plan and New Deal banking reform*. Armonk: M.E. Sharpe.
Sandilands, R. 1990. *The life and political economy of Lauchlin Currie: New dealer, presidential adviser, and development economist*. Durham: Duke University Press.
Sandilands, R. 2000. Guilt by association? Lauchlin Currie's alleged involvement with Washington economists in Soviet espionage. *History of Political Economy* 32: 473–515.
Sandilands, R., ed. 2004. New light on Lauchlin Currie's monetary economics in the New Deal and beyond. *Special Issue of Journal of Economic Studies* 31 (3/4): 170–197.
Sweezy, A. 1972. The Keynesians and government policy, 1933–1939. *American Economic Review* 62: 116–124.
Tobin, J. 1976. Hansen and public policy. *Quarterly Journal of Economics* 90: 32–37.

Customs Unions

Arthur Hazlewood

A customs union consists of two or more countries which have no tariff barriers between themselves and a common tariff against the rest of the world. There are variants which involve a greater or lesser degree of economic integration. A free trade area has no common external tariff; a union with free movement of production factors, particularly of labour, is often called a common market.

Theorizing about customs unions goes back to the classical economists, but contemporary

theory, which has been mainly concerned with the welfare effects of union, is founded on the work of Jacob Viner (1950). He introduced the concepts of trade creation and trade diversion. Trade creation occurs when the removal of the tariff on intra-union trade shifts members' demand from domestic production to lower-cost output from a union partner, diversion when the tariff preference for union members shifts demand from a non-union source to a higher-cost union supplier. The establishment of the union, according to the theory, improves welfare if trade creation predominates and worsens it if trade diversion predominates.

The removal of tariffs between union members might appear to be a move towards free trade, and presumed to be beneficial. The possibility that trade diversion will predominate shows the inadequacy of that presumption. Viner's analysis exemplifies the 'theory of the second best', showing that an incomplete move towards the optimum – the removal of some, but not all tariffs – may, in fact, make matters either better or worse.

The definition of trade creation and diversion was usefully widened by Johnson (1962) to include consumption effects. Changes in the pattern of consumption following union may either increase or decrease consumers' surplus, depending on whether there is a shift to a lower or a higher cost source of satisfying demand. The distinction between this consumption effect and the production effect of the tariff-induced price changes can be expressed as that between inter-commodity and inter-country substitutions.

Discussion stimulated by Viner's analysis was particularly concerned with establishing general conditions determining whether a union was trade creating or diverting, and the consequent welfare effects. Various assumptions were relaxed – such as that of constant costs or of fixed proportions in consumption – and various results obtained. However, no universal law of customs unions, and few conclusions of practical importance, emerged (Krauss 1972).

Meade (1955) and Johnson (1962), consistent with Viner's view that confident judgements cannot be made for customs unions in general and in the abstract, produced tentative and practically-oriented analyses of the conditions which favour a union's being on balance trade-creating, and there would be wide agreement on the following list:

1. Many union members.
2. Trade a small proportion of members' production, a high proportion giving more opportunity for a trade-diverting switch from non-union to union supply.
3. A high proportion of what trade there is being with members, and a low proportion with the outside world, again reducing trade diversion possibilities.
4. A low common external tariff as compared with the members' pre-union average tariff, further reducing the likelihood of diversion.
5. A wide overlap in the activities protected by the tariff in the different member countries, since with no overlap, there can be no trade-creating production effect through a shift in demand from a domestic to a union supplier, and a trade-diverting switch from a non-union supplier is probable.
6. Wide differences between union members in the cost of producing particular commodities.

These last two conditions provide for the countries forming the union to be actually competitive but potentially complementary.

In the light of these conditions, less-developed countries appear as most unlikely candidates for membership of a beneficial customs union. The theory deals with production and consumption shifts towards a more or a less efficient use of resources and satisfaction of consumer preferences. In developed, diversified economies such shifts can take place in response to price changes. The theory is much less adequate as an explanatory device for export-oriented, less-developed economies which produce a narrow range of commodities, and in which increased welfare requires primarily the growth and diversification of output. Yet there has been much activity in the formation of unions of such countries. Extension of the theory was required to explain this paradox, and to determine whether or not the formation of a union was economically rational: Johnson (1965);

Cooper and Massell (1965). There has, however, been little development of a theory of customs unions in the context of economic growth (see Robson 1983, ch. 2).

The basic theory acknowledges the existence of economies of scale. These may have cost-reduction effects with increased sales to union partners bringing domestic producers to a lower point on their supply curves. They may also have 'trade suppression' effects, with a switch to domestic products (in contrast with the standard trade diversion switch to products of union partners) from lower-cost (ex-tariff) non-union imports. These two effects may be thought of as parts of trade creation and trade diversion, respectively.

In a union of less-developed countries, economies of scale may have a much more central role. They may allow the development of competition between enterprises which would have monopolistic powers within the small domestic markets. Within the domestic markets many goods can be produced, if at all, only with extremely high protection. Access to the larger market allows a more efficient level of production. It also provides, above all, a stimulus to investment and economic growth. Many industries, operating with economies of scale, will be established only if they have access to the protected union market. These new industries, established to supply a demand previously satisfied entirely by imports, are by definition import-substituting and trade diverting. From this viewpoint, trade diversion becomes beneficial. In fact, it has become a major purpose of the union.

Other aspects of a customs union must be embraced by an extended theory if it is to be of particular relevance to less-developed economies. There is the existence of non-tariff barriers, which are often more serious restraints on trade than tariffs. There is the role of transport, because in many less-developed countries transport routes do not satisfactorily link potential union members, so that the removal of tariffs between them would be largely a formality, without any great effect on their trade. And there is the question of the distribution of the effects of union. For example, the location of new industries may be very unequal between members, so that some control of location or other equalizing procedures may be required. The basic theory says little or nothing about these matters, but in practice measures to deal with them are of fundamental importance to the viability of any customs union.

There is a further difficulty with the basic theory. It shows the circumstances in which the establishment of a customs union brings an improvement over protection on a national basis. However, precisely the same arguments about the efficient distribution of resources also show that free trade is better than a customs union. So why should customs unions be formed when, at their trade-creating best, they are inferior to the non-discriminatory removal of tariffs? Are customs unions simply irrational?

An explanation of why countries form customs unions, and why it can be rational for them to do so, requires the theory to be extended beyond the confines of the conventional assumption that welfare depends on private consumption alone. The inclusion of public goods, or public preferences, in the welfare function allows for policies to be counted as beneficial that would otherwise be irrational.

This approach provides a rationale for the policy commonly found in less-developed countries of attempting to secure a level of industrialization higher than would result from the operation of a free market. The preference for industry may have a non-economic basis, but it may also have economic rationality. It may be based on a belief in the importance of external economies created by industry and their beneficial effects on economic growth. In other words, the preference may have a long-run, growth-oriented basis, rather than a short-run allocation-oriented basis. The preference results in a policy of industrial protection as against free trade. Given economies of scale, this public preference for industry can be satisfied more efficiently within a customs union than within the smaller markets provided by protection on a national basis.

The device of counting as beneficial what would otherwise be seen as the opposite may be applied to a range of policies. The difficulty is that it can too easily be misinterpreted to justify any

policy; a preference that cannot be questioned may be what the governments of less-developed countries want, but not what the people need.

Customs unions were sometimes a feature of colonial arrangements, as in East Africa and Southern Africa. Many schemes have been formulated and some put into effect in the era of independence. They have not been across-the-board preferential systems, relying on the response of the market to the resulting price signals, as in the theoretical model. They have recognized the need for complex regulations if an acceptable distribution of the gains is to be achieved. There has been planned industrial specialization and location, partial protection for particular industries within the union, and fiscal redistribution. Despite these arrangements, success has been less than assured. The East African union was dissolved. In Central America the customs union, though not formally dissolved, is effectively moribund. In general, schemes failed to progress once the force of the original initiative faded.

See Also

- ▶ Economic Integration
- ▶ International Trade
- ▶ Meade, James Edward (1907–1995)
- ▶ Viner, Jacob (1892–1970)

Bibliography

Cooper, C.A., and B.F. Massell. 1965. A new look at customs union theory. *Economic Journal* 75(300): 742–747.

Corden, W.M. 1972. Economies of scale and customs union theory. *Journal of Political Economy* 80(3): 465–475.

Johnson, H.G. 1962. *Money, trade and economic growth*. London: George Allen & Unwin.

Johnson, H.G. 1965. An economic theory of protectionism, tariff bargaining, and the formation of customs unions. *Journal of Political Economy* 73(3): 256–283.

Krauss, M.B. 1972. Recent developments in customs union theory: An interpretive survey. *Journal of Economic Literature* 10(2): 413–436.

Lipsey, R.G. 1960. The theory of customs unions: A general survey. *Economic Journal* 70(279): 496–513.

Meade, J.E. 1955. *The theory of customs unions*. Amsterdam: North-Holland.

Robson, P. 1980. *The economics of international integration*. London: George Allen & Unwin.

Robson, P. 1983. *Integration, development and equity*. London: George Allen & Unwin.

Viner, J. 1950. *The customs union issue*. New York: Carnegie Endowment for International Peace.

Cycles in Socialist Economies

D.M. Nuti

In the Marxist–Leninist project of socialist economy the elimination of cycles in economic activity is the expected result of central planning replacing the ‘anarchy’ of capitalist markets. *Ex-ante* coordination of the activities of government, households and firms according to a consistent, feasible and efficient plan should, in principle, ensure the continued full employment of labour and other resources along smooth growth paths instead of the recurring bouts of booms and recessions and persistent unemployment characteristic of capitalism.

The experience of those capitalist countries which, especially since World War II, have tried to implement a social-democratic version of this project while maintaining free enterprise does not differ significantly, at least qualitatively, from that of more conventional capitalist economies. Built-in stabilizers and anticyclical management of demand may have reduced the amplitude of fluctuations and the depth of unemployment (though some government intervention has been deemed cyclical because of leads and lags); the individual cost of fluctuations and unemployment has been partly collectivized by the welfare state; but the undesired phenomena have persisted. The same is true for Yugoslavia, a country which has implemented an associationist form of socialism introducing self-management on a large scale but has retained enterprise initiative and markets.

Other countries attempted to implement the marxist-leninist project – state ownership, central

planning, equalitarianism, 'democratic centralism' under the leadership (and practical monopoly of power) of the communist party, such as the Soviet Union, the East European Six, Mongolia, China, Cuba and the other countries loosely classed as centrally planned economies or CPEs. These countries have been successful in eliminating fluctuations in the degree of labour employment. Full employment of labour was reached in the Soviet Union at the inception of the First Five-Year Plan (1928) as a result of full-scale mobilization of labour and in the other countries in the course of reconstruction after the wars that brought about the new system. Ambitious accumulation policies maintained full employment; the wage pressure generated by labour shortage itself, combined with government commitment to price stability, added sustained excess demand for consumption which contributed further to full employment stability, without any need for specific policies to support it. Full employment has been the by-product of growthmanship. In view of the persistent microeconomic inefficiency of central planning and the underfulfilment of labour productivity targets it can also be said, in a sense, that full employment of labour has been achieved 'by default'. If, however, the decentralization process currently undertaken in most centrally planned economies were to reproduce unemployment tendencies no doubt specific policies would be adopted to restore and stabilize full employment.

Outside labour employment the performance of socialist planning has been less satisfactory than originally expected. In the Soviet Union, since the completion of reconstruction and the launching of accelerated industrialization in 1928, and in the other socialist countries since the corresponding dates in their economic history, fast growth of all performance indicators in peacetime until *circa* 1960 has smoothed small-scale cyclical phenomena, reducing them to fluctuations of positive growth rates rather than of levels of income and consumption. Since then, partly because of the gradual exhaustion of labour reserves and of easily accessible natural resources, partly because of the systemic microeconomic inefficiency exacerbated by the lack of such reserves, a discernible slowdown of growth trends has been accompanied by the appearance of negative rates, i.e. fluctuations of levels as in capitalist countries. Instances range from the early minor case of Czechoslovakia in 1963 to the large-scale income drop of one third in three years in Poland 1980–82.

These phenomena are only partly attributable to exogenous shocks and their echoes, whose persistence in the socialist economy was recognized by Oskar Lange (1969), or to adjustment processes such as accelerator-type movements, whose persistence in the socialist economy had been anticipated by Aftalion already in 1909 and recognized by Notkin (1961) and Coblijc– Stojanovic (1969). Partly – indeed mostly – these phenomena are caused by systemic factors which could be classed under three groups: (i) the lack, or at any rate the slowness, of automatic adjustment feedbacks in the economic life of centrally planned economies; (ii) the acceleration of economic activity towards the end of the planning period – be it a month, a year or five years– to avoid the formal and informal penalties of underfulfilment of targets and to obtain the rewards associated with fulfilment and overfulfilment, followed by slackening at the beginning of the next period; (iii) the presence of political feedbacks, such as popular discontent and unrest resulting from deteriorating economic performance, the changes in political centralization induced by manifestations of unrest, the economic management changes associated with political changes; these phenomena adding up to a systemic mechanism of economic/political cycles.

Markets, like all servomechanisms or homeostatic (self-regulating) devices, are neither costless nor instantaneous but are automatic in their operation; at the cost of unemployment and possibly with a considerable lag, for example, an unexpected contraction in world trade can be gradually accommodated through lower wages and prices than would otherwise have prevailed, lower exchange rate and higher interest rates regardless of government intervention, capital flows etc. Central planning, like manual control, may or may not be faster and cheaper, or more

accurate, than automatic servomechanisms, depending on the relative quality of alternative controls and the actual circumstances, but is never automatic. The experience of centrally planned economies has shown repeated and sometimes glaring instances of inertia and sluggish response to exogenous change, such as persistent accelerated accumulation in the face of rising labour shortages, wage and price stability administratively enforced in spite of rising excess demand for labour and goods, systematic underpricing of imported materials and of exportables in spite of sharpening external imbalance. Reliance on monetary budget constraints and the continued presence of consumers' discretion (if not sovereignty) and some managerial room for manoeuvre make these forms of inertia and delayed response an important handicap for central planners trying to outperform market adjustments. It is precisely inadequate central response to a changing environment (including inadequate ability to innovate institutions and technology) that has given impetus to repeated attempts at reform in the last two decades.

The incentive system typical of central planning, strongly and discontinuously geared to the degree of fulfilment of physical targets, leads to frantic speeding-up of activity (*shturmovshchina* in Russian, literally 'storming') towards the end of the planning period. For monthly plans this haste leads to frequent quality deterioration; for yearly plans 'storming' leads to output being overestimated, or 'borrowed' from the subsequent period (i.e., made up through subsequent unrecorded additional output); so much so that the ratio of December output to that of the following January can be regarded as an index of economic centralization (Rostowski and Auerbach 1984). For five-year plans, 'storming' implies a concentration of investment project completions towards the end of the period and a spate of new starts at the beginning, with corresponding fluctuations. Moreover, the generalized growthmanship and emphasis on capital accumulation typical of the centrally planned economy leads usually to the inclusion in investment plans of more projects than can be completed on schedule, through 'investors' (local authorities, ministries, enterprises) underestimating true requirements in order to get a place in the plan and later escalating their demands, and through central planners systematically overestimating capacity and especially labour productivity prospects. Sometimes investment ambition leads to additional investment projects being added after or outside the plan balance (as in Gierek's Poland). As they say in East European literature, 'the investment front widens'. Sooner or later specific or generalized bottlenecks of productive or import capacity slow down implementation and reduce or block new starts. Efficiency falls due to investment resources being frozen for periods longer than economically and technically justified, and possibly because of disruption elsewhere in the economy due to resources being sucked in by investment projects given priority over current operations (a 'supply-multiplier' effect). Capital – i.e. in Marxian terminology 'dead labour' – is made unemployed instead of live labour. The cyclical pattern of starts and completions of projects, mostly within the plan period but sometimes overstepping it, leads to cyclical patterns of capacity and output endogenously generated by the system and not justified by exogenous factors. These processes have been investigated theoretically and empirically by Olivera (1960), Goldman (1964 and 1965), Baijt (1971), Bauer (1978), Dahlstedt (1981), Dallago (1982) and above all by Bauer (1982, in Hungarian, forthcoming in English).

Political factors induce cycles in socialist economy directly, through successive leaders trying to reinforce the legitimacy of their rule by appeasing their subjects with short-lived but significant spurts of consumption before the standard growth and accumulation oriented policy typical of socialist governments is resumed and comes up against the constraints discussed in the previous paragraph (Mieczkowski 1978; Hanson 1978; Bunce 1980; Lafay 1981). The association of economic and socio-political factors is investigated by Eysymontt and Maciejewski (1984), who apply discriminant analysis to a large number of indicators of such factors over time in order to identify – and anticipate – periods of crisis; they do not, however, have a model of the actual interaction of political and economic factors. An attempt at constructing such a model is made by

Nuti (1979, 1985): a critical relationship is assumed between political centralization and popular unrest, inverse up to a threshold level and direct beyond it; economic centralization is directly related to political centralization and affects – through its impact on investment policy – the level of shortages and inefficiency which in turn fuel political unrest. A recursive model with lagged variables is shown to simulate the kind of recurring rounds of reform attempts and accumulation drives observable in actual socialist economies. Screpanti (1985) has modified such a model applying catastrophe theory and obtaining a political/economic accumulation cycle similar to that of capitalist economies.

The further progress of economic reform in centrally planned economies towards market socialism is bound to attenuate and ultimately eliminate the systemic types of economic cycles discussed above. However, as Maurice Dobb had already anticipated in 1939, the diffusion of markets instead of solving the instability problems of the centrally planned economy transforms them into those typical of capitalist economies.

See Also

- ▶ Business Cycles
- ▶ Market Socialism
- ▶ Political Business Cycles
- ▶ Socialism
- ▶ Trade Cycle

Bibliography

Aftalion, A. 1909. La réalité des superproductions générales. *Revue d'Economie Politique* 23(3): 201–229.

Baijt, A. 1971. Investment cycles in European socialist economies: A review article. *Journal of Economic Literature* 9(1): 56–63.

Bauer, T. 1978. Investment cycles in planned economies. *Acta Oeconomica* 21(3): 243–260.

Bauer, T. 1982. *Tervezès, beruchàzàs, ciklusok*. Budapest: KJK.

Bunce, V. 1980. The political consumption cycle: A comparative analysis. *Soviet Studies* 32(2): 280–290.

Coblijc, N., and L. Stojanovic. 1969. *The theory of economic cycles in a socialist economy*. New York: IASP.

Dallago, B. 1982. *Sviluppo e Cicli nelle Economie Est-Europee*. Milan: Angeli.

Dobb, M.H. 1939. A note on saving and investment in a socialist economy. *Economic Journal* 43: 713–728.

Eysmontt, J. and Maciejewski, W. 1984. Kryzysy spolecznogospodarcze w Polsce – ujecie modelowe (Social-economic crisis in Poland-a model approach). *Ekonomista.*

Goldmann, J. 1964. Fluctuations and trends in the rate of economic growth in some socialist countries. *Economics of Planning* 4(2): 88–98.

Goldmann, J. 1965. Short and long term variations in the growth rate and the model of functioning of a socialist economy. *Czechoslovak Economic Papers* 5: 35–46.

Hanson, P. 1978. Mieczkowski on consumption and politics: A comment. *Soviet Studies* 30(4): 553–556.

Lafay, J.-D. 1981. Empirical analysis of politico-economic interaction in East European countries. *Soviet Studies* 33(3): 386–400.

Lange, O. 1969. *Theory of Reproduction and Accumulation*. Oxford: Pergamon.

Mieczkowski, B. 1978. The relationship between changes in consumption and politics in Poland. *Soviet Studies* 30(2): 262–269.

Notkin, A. 1961. *Tempy i proportsii sotsialisticheskogo vosproizvodstva (The rate and proportions of socialist reproduction)*. Moscow: IEL.

Nuti, D.M. 1979. The contradictions of socialist economies: A Marxian interpretation. *Socialist Register*. London: The Merlin Press.

Nuti, D.M. 1985. *Political and economic fluctuations in the socialist system*, Working Paper No.85/156. Florence: European University Institute.

Olivera, J. 1960. Cyclical growth under collectivism. *Kyklos* 13(2): 229–252.

Rostowski, J. and Auerbach, P. 1984. Storming cycles and central planning. Discussion Paper in Political Economy No.52, Kingston Polytechnic.

Screpanti, E. 1985. *A model of the political economic cycle in centrally planned economies*, Working Paper No.85/201. Florence: European University Institute.

Cyclical Markups

Julio J. Rotemberg

Abstract

This article first shows that countercyclical variations in the ratios of prices to marginal cost (markups) can cause pro-cyclical fluctuations in the demand for labour at a given real wage and thus induce fluctuations in economic

activity that look like business cycles. It then discusses methods for measuring cyclical movements in markups and shows that several types of evidence suggest that these are counter-cyclical. Lastly, it discusses economic mechanisms that can explain these counter-cyclical markup movements.

Keywords

Cobb–Douglas functions; Cyclical markups; Elasticity; Imperfect competition; Increasing returns; Inventory investment; Labour productivity; Labour supply; Leisure; Limit pricing; Real business cycles; Sticky prices; Wealth effects

JEL Classification

D4; D10

Firms that have increasing returns to scale, that produce differentiated products, or that are part of a small oligopoly can generally be expected to set a price above marginal cost. In so far as a firm's ratio of price to marginal cost is larger than one, there is no particular reason to suppose that this ratio, or markup, will stay constant when overall economic conditions change. Indeed, different models of imperfect competition have different predictions concerning how this markup should vary as aggregate income and activity expands and contracts. Thus, an analysis of whether markups rise when aggregate activity rises or whether they rise when aggregate activity declines provides a useful lens for determining which theories of firm behaviour have more validity.

Markup variations are also of central importance for macroeconomics. One of the central questions for macroeconomics is why the economy expands and contracts at cyclical frequencies in the first place, and cyclical movements in markups are potentially an important nexus that allows such fluctuations to occur. When a single firm (or industry) raises the ratio of its price to its marginal cost, one expects its relative price to rise so that the quantity it sells falls. However, when every firm in the economy tries to raise its price relative to its marginal cost, relative prices need not be affected.

When every firm raises its markup two important consequences follow. The first is that real marginal cost, which can be defined as nominal marginal cost divided by the typical price charged by firms, must fall. Thus, the question of whether markups are countercyclical is the same as the question of whether real marginal costs are pro-cyclical. The second consequence of all firms varying their markups at the same time is that the aggregate demand for labour changes. To see this, notice that nominal marginal cost is equal to the nominal wage divided by the marginal product of labour. Thus, a generalized increase in markups means that prices must rise relative to nominal wages if employment is to remain at a level that keeps the marginal product of labour constant. Alternatively, firm are willing to pay the same real wage only if the marginal product of labour rises, and this requires that employment fall if labour is subject to diminishing returns. In either way of seeing this change, the demand for labour at any given wage falls.

The Role of Markup Changes in Economic Fluctuations

The capacity of markup changes to generate changes in aggregate labour demand is important because several pieces of evidence suggest that short-run business fluctuations are the result of changes in the demand for labour. That the willingness of firms to hire labour at any given real wage increases in economic expansions is suggested first of all by the tendency of real wages to increase when the economy expands. As shown by Bils (1985), this tendency is particularly strong when one looks at the wages of individuals (as opposed to looking at average wages paid to all workers). Moreover, as emphasized by Bils (1987), firms tend to use more overtime hours in economic booms, and firms are legally obliged to pay higher hourly wages for these overtime hours. When combined with the pro-cyclicality of real wages, other pieces of

evidence also suggest that labour demand is higher in booms. In booms, both the unemployment rate and the fraction of the unemployed who have been unemployed for longer than 5 weeks tend to be lower (both of which suggest that finding jobs is easier) and that the number of help-wanted advertisements is larger (suggesting that it is more difficult for firms to find workers even as they pay them higher wages).

The real business cycle literature stresses a different source of labour demand movements: namely, exogenous changes in the productivity of the typical firm. This hypothesis has the advantage that it explains in a straightforward fashion why labour productivity is somewhat pro-cyclical. However, as discussed below, movements in markups lead to pro-cyclical productivity under a variety of plausible assumptions. In this regard, a clear advantage of the view that markup movements are responsible for important labour demand movements is that labour productivity and real wages rise together with output also when output increases appear to be due to non-technological factors such as increases in military spending, expansionary monetary policy or reductions in the price of oil. (Evidence of these conditional correlations of productivity and output can be found in Hall 1988.)

Relative to markup variations, exogenous short-run changes in technical progress have another disadvantage as sources of cyclical fluctuations. This is that technical progress not only increases the willingness of firms to hire workers but also reduces the willingness of workers to work at any given wage. These contractionary movements in labour supply are the result of 'wealth effects': technical progress makes people richer and thus induces them to consume both more goods and more leisure. These effects are particularly large if technical progress is somewhat permanent, as tends to be true with actual examples of such progress. These reductions in labour supply imply that shocks to technical progress have only small expansionary effects on employment. By contrast, reductions in markups induce only modest wealth effects, so employment responds more strongly to the resulting increases in labour demand.

These conceptual benefits of countercyclical markups raise the question of whether markups do indeed rise in economic contractions and fall in booms. To discuss this, it is worth starting with the case where the value added production function takes the Cobb–Douglas form. With capital essentially fixed in the short run, this implies that aggregate value added Y is equal to the labour input H to the power α. The marginal product of labour is then equal to α times the average product of labor Y/H The ratio of marginal cost to price is then the wage divided by both the marginal product of labour and the price, so that it is proportional to the labour share in value added (or unit labour cost) WH/PY.

Measuring Markup Variations

If aggregate data are used, the labour share in value added is not a very cyclical variable. Labour productivity Y/H tends to rise mildly in expansions, as does the average real wage – though the size of these effects depends on how one measures economic expansions. Because cyclical productivity changes are slightly larger than the corresponding average changes in real wages, the labour share has a modest tendency to fall in expansions. If the labour share were seen as equal to the inverse of the markup (as implied by the Cobb–Douglas assumptions), markups would be pro-cyclical and actually dampen cyclical fluctuations.

As suggested in the survey by Rotemberg and Woodford (1999), this Cobb–Douglas case is a good baseline, but a number of corrections to the resulting measure of the markup immediately suggest themselves, and these tend to make measured markups more counter-cyclical. The first of these is that, as already alluded to above, what matters for marginal cost is not the average wage but the marginal wage for an additional hour of work. The average wage is dragged down in booms by the absorption into employment of many relatively low-wage workers who are not employed in recessions. If these workers are less productive, their wage per effective unit of labour input may actually be relatively large. Whatever the case, individual workers who remain employed do see their

wages rise more substantially, as emphasized by Bils (1987). Admittedly, these wage increases are concentrated among workers who change jobs, and the increases in the 'straight-time' wages of people who stay in the same job are more modest. The marginal hour of work, on the other hand, is more likely to be an overtime hour in booms, and this is probably the most important reason for believing that the marginal hour of labour is more expensive then.

It also seems important to correct the way the Cobb–Douglas approach measures the marginal product of labour. According to this functional form. the marginal product of labour is simply proportional to the average product of labour. Given that the average product of labour actually rises slightly in booms, this functional form essentially requires that the economy become 'more productive' in booms, perhaps as a result of increased technical progress.

The tendency of labour productivity to be pro-cyclical can be interpreted in two rather different ways, both of which have a direct bearing on calculations of the cyclical properties of the marginal product of labour. The first is that firms are subject to increasing returns to scale. The simplest functional form that captures this supposes that there are fixed costs, that is, that some of their inputs are 'overhead' inputs that are required to produce even a minuscule positive quantity of output for sale. Suppose for example, that $\overline{H}$ units of labour are overhead units so that output continues to be given by the Cobb–Douglas form but is now proportional to $(H-\overline{H})$ to the power α. The marginal product of labour is then proportional to the ratio $Y/(H-\overline{H})$. In booms, the percentage increase in $H-\overline{H}$ obviously exceeds the percentage by which H rises so that $Y/(H-\overline{H})$ falls by more than Y/H. This means that for $\overline{H}$ sufficiently large, the marginal product of labour falls, marginal costs rise and measured markups fall. Assuming that some of the labour input takes this overhead form can thus easily lead to the inference that markups are indeed counter-cyclical.

A second possible reason for the observation that the average product of labour is pro-cyclical is that firms do not fully utilize all their labour in recessions. They 'hoard' labour to avoid having to incur hiring and training costs when economic activity recovers. This raises two important questions. The first is whether workers produce something else other than measured output when they are being hoarded. The second is whether the firm needs to pay them less when their GDP-producing effort is lower. Given that real wages are only slightly pro-cyclical, it is probably more realistic to suppose that the cost of an hour of labour services to the firm is the same whether the worker incurs effort (and produces) or not. Particularly if the workers are not producing much unmeasured output in recessions, this implies that marginal cost in recessions is considerable smaller than is implied by H/Y. Real marginal cost is more pro-cyclical than WH/PY and markups are more counter-cyclical. One attractive feature of this explanation for pro-cyclical labour productivity is that it is very compatible with the idea that markups are counter-cyclical. Firms are willing to keep workers idle in recessions even though marginal cost is extremely low precisely because they are keeping their prices high relative to marginal cost.

There are two additional types of evidence suggesting that markups are relatively low in booms and high in recessions. The first comes from the behaviour of intermediate inputs relative to final goods. A crude view of materials is that these are used in fixed proportions relative to the gross output of final goods. However, Basu (1995) shows that the ratio of materials to final goods tends to rise when the economy expands. If the material intensity of output is a choice variable, the ratio of marginal cost to price must also equal the real price of materials divided by the marginal product of materials. It is reasonable to suppose with Basu (1995) that the marginal product of materials diminishes as the level of materials inputs rises. With constant returns, the increase in the ratio of materials to output in booms thus implies that real marginal cost is pro-cyclical even if the price of materials relative to final output were constant. In fact, Murphy et al. (1989) show that that prices of more processed goods tend to fall relative to prices of less

processed goods in economic expansions, and this too indicates a tendency of price to fall relative to marginal cost during booms.

The second additional source of evidence comes from the behaviour of inventories. Inventories rise in booms but, as stressed by Bils and Kahn (2000), they rise by less in percentage terms than sales. At the same time, long-run growth in sales does tend to be associated with equiproportonate increases in inventories in the industries they consider. In addition, they discuss cross-sectional evidence that shows that, within industries, the inventory–sales ratios of products with large sales are not smaller than the inventory-sales ratios for products with low sales. This suggests that there is something special about the decline in inventory–sales ratios that is observed in booms. It suggests, in particular, that conditions in booms lead firms to economize on inventory holding. As Bils and Kahn (2000) argue, the evidence seems most consistent with the idea that firms keep their inventories relatively low in booms because real marginal cost is relatively high.

Theories of Cyclical Markup Variations

A considerable body of evidence, then, seems consistent with counter-cyclical markups, and suggests that countercyclical markups might be central to aggregate fluctuations because they rationalize the changes in employment that characterize such fluctuations. The question that remains is why markups should vary cyclically. There are basically five types of models that explain these movements in markups. These are: models of variable demand elasticity, models of variable entry, models of sticky prices, models of investment in market share and models of implicit collusion.

In a monopolistically competitive setting, markups are equal to the elasticity of demand over the the elasticity of demand minus 1. Increases in the elasticity of demand thus lower markups (towards the competitive level of 1) and could thus be a source of business expansions. This still leaves the question of why the elasticity of demand facing the typical firm should vary over time. One possibility is that the proportion of demand that comes from highly elastic customers rises in booms. Gali (1994) obtains such composition effects under the supposition that investment is more price sensitive than consumption. Ravn et al. (2004) obtain a related effect by supposing that people have formed a 'habit' for at least a fraction of past purchases, and the elasticity of demand for these habitual purchases is negligible relative to the elasticity of demand for non-habitual ones. As consumption rises in economic expansions, more of the purchases are non-habitual so that the elasticity of demand is higher and markups have to be correspondingly lower.

Devereux et al. (1996) show that changes in demand induced, for example, by changes in government purchases lead new firms to enter existing industries. Entry of new firms is indeed quite pro-cyclical. Such entry can, in turn, make each firm's perceived elasticity of demand higher (because they fear more competitors). Thus variable entry can be seen as a reason for changes in elasticities that lead to counter-cyclical markups. Even if the expansion in the number of firms that takes place in booms is seen as too small for this effect to be large, the potential for increases in entry may lead incumbents to keep their prices low to avert the creation of an even larger number of new firms. This limit pricing might also be able to rationalize counter-cyclical markups.

Sticky prices, which are widely assumed in new Keynesian macroeconomics, probably provide the most straightforward model of counter-cyclical markups. Firms that keep their prices constant when demand increases (as a result of expansionary government policy, for example) will generally see their marginal costs rise both because of diminishing returns and because of increases in the costs of factor inputs. Thus, keeping their prices relatively constant will lead them to have lower markups. The argument that sticky prices derive their influence on the economy from their consequences for variable markups is presented in more detail in Kimball (1995).

If customers who have already purchased a good have relatively inelastic demand, keeping price low is like an investment activity for the firm. It encourages new customers (those whose demand is elastic) to become addicted. Changes in

economic conditions can lead firms to desire to either increase or decrease these investments. Increases in interest rates in particular might lead firms to wish to reduce these investments, at least temporarily. Chevalier and Scharfstein (1996) provide evidence that the cash condition of firms plays a large role in these investments as well. They show that recessions have a disproportionate effect on the pricing of cash-strapped firms, who turn out to be more eager to raise prices and thereby reduce their investment in market share.

Lastly, Rotemberg and Saloner (1986) have emphasized that high prices may be more difficult to sustain for implicitly collusive oligopolists in economic expansions. When current sales are high, each firm perceives a greater benefit from undercutting the implicit agreement because it can thereby secure even higher sales. To prevent this, the oligopolists must lower their markups of price relative to marginal cost. Some cross-sectional evidence suggests that markups are indeed more counter-cyclical in more concentrated sectors, as a theory that applies only to implicitly collusive oligopolists suggests. As shown by Rotemberg and Woodford (1992), the model can be embedded in a general equilibrium structure so that increases in government purchases raise output together with real wages. The increased rate of interest induced by additional government purchases lowers the present value of the future benefits from cooperation. It thus forces oligopolies to be less ambitious in the profits that they seek from current prices, so that markups fall and labour demand rises.

See Also

- ▶ Microfoundations
- ▶ New Keynesian Macroeconomics

Bibliography

Basu, S. 1995. Intermediate inputs and business cycles: Implications for productivity and welfare. *American Economic Review* 85: 512–531.

Bils, M.J. 1985. Real wages over the business cycle: Evidence from panel data. *Journal of Political Economy* 93: 666–689.

Bils, M.J. 1987. The cyclical behavior of marginal cost and price. *American Economic Review* 77: 838–857.

Bils, M., and J.A. Kahn. 2000. What inventory behavior tells us about business cycles. *American Economic Review* 90: 458–481.

Chevalier, J.A., and D.S. Scharfstein. 1996. Capital-market imperfections and countercyclical markups: Theory and evidence. *American Economic Review* 86: 703–725.

Devereux, M.B., A.C. Head, and B.J. Lapham. 1996. Monopolistic competition, increasing returns, and the effects of government spending. *Journal of Money, Credit, and Banking* 28: 233–254.

Gali, J. 1994. Monopolistic competition, business cycles, and the competition of aggregate demand. *Journal of Economic Theory* 63: 73–96.

Hall, R.E. 1988. The relation between price and marginal cost in U.S. industry. *Journal of Political Economy* 96: 921–947.

Kimball, M.S. 1995. The quantitative analytics of the basic neomonetarist model. *Journal of Money, Credit, and Banking* 27: 1241–1277.

Murphy, K.M., A. Shleifer, and R.W. Vishny. 1989. Building blocks of market clearing business cycle models. In *NBER macroeconomics annual*, ed. O.J. Blanchard and S. Fischer. Cambridge, MA: MIT Press.

Ravn, M., S. Schmitt-Grohe, and M. Uribe. 2004. *Deep habits*, Working Paper No. 10261. Cambridge, MA: NBER.

Rotemberg, J.J., and G. Saloner. 1986. A supergame-theoretic model of price wars during booms. *American Economic Review* 76: 390–407.

Rotemberg, J.J., and M. Woodford. 1992. Oligopolistic pricing and the effects of aggregate demand on economic activity. *Journal of Political Economy* 100: 1153–1207.

Rotemberg, J.J., and M. Woodford. 1999. The cyclical behavior of prices and costs. In *Handbook of macroeconomics*, vol. 1B, ed. J.B. Taylor and M. Woodford. Amsterdam/New York/Oxford: North-Holland.